Humboldt-Universität zu Berlin

Institut für Bibliotheks- und Informationswissenschaft

Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft

Heft 470

Quantitative assessment of metadata collections of research data repositories

von
Dorothea Strecker

Quantitative assessment of metadata collections of research data repositories

von

Dorothea Strecker

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Vivien Petras
Humboldt-Universität zu Berlin

Heft 470

Strecker, Dorothea

Quantitative assessment of metadata collections of research data repositories / von Dorothea Strecker. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2021. – 68 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 470)

ISSN 14 38-76 62

Abstract

Structured metadata are of particular importance in the context of facilitating research data (re-)use. Although research data repositories create and manage metadata records, existing research offers limited insights into the relationship between repositories and metadata for research data.
Therefore, in conducting a quantitative assessment informed by metadata quality requirements, this thesis aims at making distinctive features of metadata for research data visible, specifying the potential influence of repository characteristics on metadata, and exploring changes to metadata records.
The analysis showed variations in metadata completeness across repositories. Within repositories, metadata descriptions are relatively homogenous. These findings suggest that repositories have developed distinctive and consistent practices for describing data. On average, descriptions comprise 487.3 characters, and 5.52 years passed between the year a dataset was published and the metadata record was registered. Differences in the completeness of metadata records, description length and timeliness were significant across repository types and certification status, whereas differences in collection homogeneity were not significant. Overall, most metadata records in the sample were changed, which conforms with the conceptualization of metadata for research data as dynamic and changeable objects. Differences in the number of changes are significant across repository types.

Diese Veröffentlichung geht zurück auf eine Masterarbeit im Studiengang Information Science, M. A. an der Humboldt-Universität zu Berlin.

Eine Online-Version ist auf dem edoc Publikationsserver der Humboldt-Universität zu Berlin verfügbar.

Contents

List of Figures

List of Tables

1 Introduction

In the context of the Open Science movement, research data are increasingly regarded as distinct and valuable research outputs. Various stakeholders argue that the value of data increases with their use and advocate the sharing and publication of research data to ensure transparency of results. There is also a growing demand for data in many fields of research. Increasingly, this need exceeds data collected by researchers themselves on the basis of their own research questions to include data recorded by other researchers, sometimes for entirely different purposes.
A broad political and cultural shift towards data sharing is further exemplified by the growing recognition of the FAIR Principles, a set of principles intended to make research data findable, accessible, interoperable and reuseable for machines and humans (Wilkinson et al., 2016). However, data sharing is not a trivial task. Before a dataset can be reused, it must cross temporal, spatial or even subject boundaries. Overcoming these boundaries to successfully share datasets often requires comprehensive descriptions, for example in the form of structured metadata records. Therefore, the FAIR Principles not only apply to datasets, but also extend to the metadata describing them (Research Data Alliance FAIR Data Maturity Model Working Group, 2020).
Research data repositories are specialized infrastructures and play a central role in data stewardship and the creation, maintenance and dissemination of metadata records for research data. The landscape of research data repositories is diverse, with variations in disciplinary focus, available resources and level of curation (Kindling et al., 2017). Collectively, these repositories cover a wide range of requirements and use cases. While there are ambitious policies in place and the number of repositories increases, little is known about the status quo and results of data stewardship (York, Gutmann, and Berman, 2018). One area where understanding is significantly lacking is metadata creation and maintenance, especially differences across repositories (Gregg et al., 2019).
Although metadata are necessary to move datasets beyond the context of their collection, the current state of metadata collections at individual research data repositories as well as the influence of repository characteristics on these collections has not been studied in detail. Another open question is whether metadata should rather be regarded as finished products or whether data descriptions should be considered an iterative process (Edwards et al., 2011).
Therefore, this thesis will investigate the influence of repositories on metadata collections, focusing on research data repositories using the DataCite Metadata Schema.

1.1 Research questions

RQ 1 What properties characterize metadata (collections) at research data repositories using the DataCite Metadata Schema?

RQ 2 Is there a relationship between characteristics of repositories and properties of their metadata (collections)?

RQ 3 Is there a relationship between characteristics of repositories and the number of changes to metadata records?

This research will contribute to existing studies of metadata for research data, focusing on a quantitative description of the relationship between repositories and their metadata collections. In addressing this gap in the literature, this thesis makes the results of metadata labor in the context of data stewardship visible and reflects them from a perspective of aspects of metadata quality.

2 Literature review

2.1 Research data

The concept *data* is often taken as self-evident and rarely questioned. However, a definition of the concept and its application in research is necessary before examining how data are handled in information infrastructures. Therefore, the starting point for this thesis is a discussion of the concept in the context of scientific knowledge production.

2.1.1 Brief history of data in knowledge production

Literally, the term *data* is the plural form of the Latin *datum*, a "(thing) given" (*Online Etymology Dictionary: data*). According to historian of information Daniel Rosenberg, the concept as we understand it today first appeared in the English language in the seventeenth century (Rosenberg, 2013). The term *data* was mainly used in the context of arguments in mathematics and theology, where it described shared principles underlying arguments or facts documented in scripture. Throughout the eighteenth century, this meaning shifted: "By the end of the century, the term was most commonly used to refer to facts in evidence determined by experiment, experience, or collection." (Rosenberg, 2013; p. 33) Rosenberg concludes that the concept *data* as we understand and use it today is relatively new.

In recent years, data have become central to knowledge production. The influential essay collection "The Fourth Paradigm" published in 2009 describes how increasing computing power and data volumes transform scientific practice and methodology in many disciplines (Hey, Tansley, and Tolle, 2009). The central idea of the book is the shift from theory-based research to inductive inference from data. Now, more than ten years after the book's publication, this paradigm shift has become apparent through the proliferation of the term *Big Data*, which implies that a knowledge production, at least in some disciplines, is based on large amounts (*volume*) of diverse data (*variety*), often complete databases (*exhaustivity*), using increasing computing power (*velocity*) (Kitchin and McArdle, 2016).

2.1.2 Definitions of research data

In information science, information is often defined in terms of interrelated concepts such as data and knowledge. The concepts and their interrelations are often modeled as a hierarchy with the levels data, information and knowledge, as well as the transitions of one stage to the next. This idea was introduced by Russell Ackoff in 1989. In Ackoff's model, data are symbols representing characteristics of objects or events, whereas Information is data processed with the intention to make it useful (Ackoff, 1989). This and similar models imply that data are objective representations of reality and lack context or meaning (Rowley, 2007).

This perspective on research data, which philosopher of science Sabina Leonelli calls the *representational view*, is also common in science: "The *representational view* construes data as reliable representations of reality which are produced via the interaction between humans and the world." (Leonelli, 2020b) The representational view maintains that through direct representation, data enable unmediated knowledge of the phenomena under investigation.
This perspective culminates in the idea of *raw data* that are foundational or *simply are*, and represent facts objectively and truthfully, independent of their context. The notion of *raw* data is challenged in disciplines like information science and philosophy of science. In the introduction to the book "Raw Data is an Oxymoron", Lisa Gitelman and Virginia Jackson argue that data are never raw, because they need to be understood in context (Gitelman and Jackson, 2013). Data are shaped by and shape the conditions of their creation and use: observations are based on the interpretation and application of theories, therefore observations (and by extension the research data generated by these observations) are already "theory-laden" (Chang, 2005; p. 882).
Leonelli proposes an alternative perspective on data that takes these objections into account - the *relational view* (Leonelli, 2020b). She emphasizes the role of research data in scholarly communication, reconciling the *prospective usefulness as evidence* with the *portability* of data (Leonelli, 2015 ; Leonelli, 2020a). Taking a relational perspective on research data means considering research data first of all as material objects that have certain characteristics, such as a format and medium, but no inherent truth value. These objects become *data* if they are treated as evidence for phenomena under investigation (*prospective usefullness as evidence*) and can be shared among researchers (*portability*). According to Leonelli's relational view, data therefore can be viewed as "[..] a relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved." (Leonelli, 2015; p. 2) Data are not evidence of phenomena in and of themselves, but become evidence by being used or interpreted in certain ways (Leonelli, 2020a). Whether digital files or other entities are considered *research data* is therefore context-dependent. The term is an attribution that defines data as "[...] the first transformation of nature in the production chain that culminates in knowledge." (Strasser and Edwards, 2017; p. 330)
In her book "Big Data, Little Data, No Data", Christine Borgman comes to a very similar conclusion after discussing existing definitions of research data (Borgman, 2016). Borgman argues that so far, there is no consensus on a definition. This lack of agreement is, among other aspects, attributable to the fact that characteristics of research data can vary notably, for example in terms of their materiality or degree of aggregation. Borgman concludes that therefore, research data are not defined by a set of specific characteristics, but by their role in the research process: "The most inclusive summary is to say that data are representations of observations, objects or other entities used as evidence of phenomena for the purpose of research or scholarship." (Borgman, 2016; p. 28)
The activity of *data collection* can also contribute to understanding research data. Bruno Strasser and Paul Edwards maintain that the activity of collecting establishes relationships between objects, for example between rock samples, their digital representations, and other collected objects. These activities are central to knowledge production, as they create a representation of nature for the purpose of research, a "second nature". (Strasser and Edwards, 2017; p. 331)
These discussions on approaches to defining *data* show that it is not a trivial or self-evident concept. It has also been shown that there are justified objections to a representational view of data. Therefore, this thesis adopts a relational perspective on research data, which are understood as representations of phenomena under investigation, and their evidential value arises from being used as evidence in a specific context.

2.1.3 Research data and Open Science

As discussed above, data are increasingly viewed as valuable assets driving scientific discovery. This commodification of data creates tensions and can lead to very different patterns of data ownership. For example, in the context of genome data, private companies are currently building opaque business models on data obtained by genetic sequencing and interpretation services for individuals (Leonelli, 2019). In contrast, the Bermuda Principles of 1996 mandate the sharing of genome sequence data in science, effectively creating a "genome commons". (Contreras, 2010; p. 63) Within this mode of data ownership, the value of data increases with its reuse, because the overall usefulness of a dataset relative to the cost of its collection grows (Palmer, Weber, and Cragin, 2011).
The Open Science movement aims at breaking imbalanced patterns in the access to knowledge. Open Science is an umbrella term covering several "schools of thought" (Fecher and Friesike, 2014). These schools of thought differ in their specific aims or methods, but they all challenge traditional modes of scholarly communication, for example by drawing attention to data, software, and other products of the research process. One area of activity is the promotion of data sharing as opposed to other modes of ownership. Various measures have been implemented to encourage researchers to adopt data sharing practices (Kim and Stanton, 2016). There are technical aspects to making research data available, but setting the right incentives for individual researchers is widely considered more challenging (Klump, 2017).
The Open Science movement therefore pursues a cultural shift in scientific communication to overcome concentrated ownership of knowledge.

2.2 Infrastructures for research data

Technical and social aspects of sharing data converge in specialized infrastructures, which will be discussed in the following.

2.2.1 Information infrastructures

Geoffrey Bowker and his co-authors define infrastructures as collective facilities, practices or standards enabling human activities, as "[...] pervasive enabling resources in network form [...]." (Bowker et al., 2010; p. 98) Infrastructures permeate different areas of life, and they can be distinguished by the activities they enable and the communities they support – in the case of information infrastructures, they enable knowledge work, including research (Bowker et al., 2010). According to Susan Leigh Star and Karen Ruhleder, infrastructures are so pervasive that they are often taken for granted and therefore fade into the background, and by implementing standards, they can be layered or connected (Star and Ruhleder, 1996).
In his Book "A vast machine", Paul N. Edwards defines information infrastructures as "[...] robust networks of people, artifacts and institutions that generate, share, and maintain specific knowledge about the human and natural worlds." (Edwards, 2013; p. 17) Within the community they serve, these systems are ubiquitous, widely accessible, standardized and reliable. They span time, space and social context, thereby providing stability for knowledge work. This makes them seem so natural that they mostly remain invisible, unless they fail (Star and Ruhleder, 1996). Information infrastructures not just comprise tangible hardware, but also persons and practices (Edwards, 2003). Edwards emphasizes the sociotechnical nature of information infrastructures – they shape and are shaped by their users, and are often a gateway to belonging to a certain community: "Belonging to a given culture

means, in part, having fluency in its infrastructures.“ (Edwards, 2003; p. 189)
Research data are an important driver in the development of information infrastructures. For example, the emergence of databases in various scientific disciplines and changes in scholarly communication (particularly digitization) required new equipment capable of handling new types and larger amounts of data (Bowker et al., 2010). Infrastructures also provide standardization, thereby reducing *data friction*: the costs associated with collecting, using, and storing data (Edwards, 2013; p. 84).
One type of information infrastructure specialized on handling research data are research data repositories.

2.2.2 Research data repositories

According to the CASRAI Research Data Management Glossary, “Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats.“(*CASRAI Research Data Management Glossary: Repository*) Research data repositories in particular focus their activities on the collection, curation, preservation and dissemination of research data (Assante et al., 2016). It is important to note that research data repositories exceed the scope of databases, as their activities are not just centered around data storage. Areas of activities can include, for example, education of researchers, implementation of policies, or taking steps to understanding data curation needs (Lee and Stvilia, 2017).
Although most research data repositories share similar objectives, specific characteristics of individual repositories and the practices they adopt to support data publishing vary across institutions and disciplines. As specialized infrastructures, research data repositories reflect the requirements of the community they serve. For example, repositories differ in terms of the content types they hold, restrictions placed on data upload or download, and the use of standards (Kindling et al., 2017). The types of services offered by a repository depend on several factors, including the software and tools used, policies and norms put in place, as well as repository staff and their skill sets (Lee and Stvilia, 2017). A survey among North American academic libraries showed that in 2014, institutional repositories were more likely to offer informational and consultative services than technical services (Tenopir et al., 2015). This lack of technical services may be related to the criticism that research data repositories tend to follow traditional paths of publishing scientific texts, and sometimes fail to take into account specific needs associated with research data, such as solutions for collaboration and handling dynamic and changing objects (Parsons and Fox, 2013; Assante et al., 2016). Although repositories differ in the services they offer, the assignment of persistent identifiers is considered an essential service by most stakeholders involved (Schwardmann, 2020). Persistent identifiers facilitate the unique and reliable identification and citation of datasets (Klump, Huber, and Diepenbroek, 2015).
Research data management underwent significant professionalization in recent years, which is reflected, for example, in an emphasis on trustworthiness, the publication of requirements and the development of formal certification initiatives. Trustworthiness of research data repositories is highlighted, since they should, for example, encourage trust among researchers who may still have reservations against data sharing (Klump, 2017). Researchers' trust in a research data repository is influenced by several factors, including the validity and accuracy of datasets in its collection and the documentation processes it adopts (Yoon, 2014). Organizational aspects also factor into trustworthiness, for example the ability to preserve datasets long-term. Maintaining data infrastructures over long periods of time is still a big challenge, however, particularly with regard to sustained funding and organizational support (Imker, 2020). Published requirements provide orientation for repositories. One example of a publication intended to shape the development of research data repositories are the TRUST Principles, formulated in non-technical terms to facilitate communication with diverse stakeholders (Lin et al.,

2020). TRUST stands for transparency, responsibility, user focus, sustainability and technology. The TRUST Principles provide guidance for the operation and development of research data repositories at a very high level of abstraction. In contrast to non-committal recommendations like the TRUST Principles, certification offers a more specific and formalized approach to ensuring adherence to a set of criteria. One of the most common certificates issued to research data repositories is the CoreTrustSeal (CTS). The CTS certification process is based on repositories' self-assessment, which is then verified by reviewers selected from the CTS Board and representatives from certified repositories (CoreTrustSeal Standards and Certification Board, 2019). CTS focuses on the trustworthiness of a research data repository, promoting confidence in using it to publish or archive datasets. Currently, the CTS comprises 16 requirements, ranging from organizational prerequisites to research data management and technology.

As mentioned in the section above, infrastructure development and use are interrelated: practices of scientific communities shape and are shaped by infrastructures (Leonelli, 2020a). This also applies to research data repositories, which are often categorized by the scope of communities they serve. In particular, the degree of disciplinary focus is a key factor in repository typology: it determines whether a repository is considered a *specialist* or *generalist* repository (Assante et al., 2016; Lee and Stvilia, 2017).

The interrelation of infrastructure development and the community it serves is the foundation of the OAIS reference model, an ISO-standard for preservation infrastructures. Within the model, the *designated community* denotes "[...] potential Consumers who should be able to understand a particular set of information." (CCSDS Secretariat, 2012; p. 1-11) Preservation infrastructures define their designated community and adopt suitable technologies and practices on this basis. Designated communities can comprise multiple (sub-)groups and change over time. If a repository's designated community shifts, changes in technology and services are likely to follow (Donaldson, Zegler-Poleska, and Yarmey, 2020). However, the concept of designated communities may not apply to repositories with broad user bases. For example, research data repositories with a mission to offer services to a wide spectrum of scientific disciplines may find it difficult to clearly define a designated community (Bettivia, 2016). Generalist repositories, for example repositories serving a research institution, report unique challenges associated with a diverse set of users (Joo, Hofman, and Kim, 2019). In addition, it is important to note that repositories also serve potential data reusers. Considering data reusers' needs requires flexibility, as not all possible scenarios of data reuse can be anticipated in advance, and datasets can be used for purposes not intended by data providers or repository staff (Parsons and Duerr, 2006; Leonelli, 2020a). For example, Late and Kekäläinen showed that data from a Finnish social science data archive was also used by researchers associated with natural, medical, and technical sciences, as well as the humanities (Late and Kekäläinen, 2020). Scientific disciplines are also not necessarily homogenous, and data practices of community members may vary notably (Mayernik, 2015). The more flexible concept of *data communities*, "[...] formal or informal groups of scholars who share a certain type of data with each other, regardless of disciplinary boundaries", may be better suited for developing infrastructures that bridge gaps between data providers and data reusers (Springer and Cooper, 2020; p. 2).

In summary, information infrastructures and the community they serve are closely interrelated. However, the community is not static and cannot be clearly separated from non-members. Repositories therefore should remain flexible with regards to anticipating the needs of their community.

2.3 Metadata for research data

Infrastructures are enablers of certain actions and practices. In the case of information infrastructures, *enabling* often involves infrastructure functions that are based on metadata.

2.3.1 Metadata

Metadata are often defined as *data about data*, based on a literal translation of the term (Pomerantz, 2015). This definition, although common, does not offer any insight into what exactly constitutes metadata, what they do, and why it matters.
Zeng and Qin describe metadata as "[encapsulating] the information that describes any information-bearing entity." (Zeng and Qin, 2016; p. 11) A narrower understanding of metadata is traditionally adopted by memory institutions such as libraries. They create detailed and highly structured metadata to manage their collections of information resources and make them accessible to users (Riley and National Information Standards Organization (U.S.), 2017; Zeng and Qin, 2016). Digital information infrastructures highlight the importance of metadata in managing and structuring collections. In this context, the term *metadata* has been used since the 1990s in reference to "[...] internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system." (Gilliland, 2008) Metadata enable key functions of digital information infrastructures, as they provide a simplified means of interacting with digital content.
This is possible because metadata take the form of statements about the object to be described, allowing for a simplified representation of that object. Consequently, metadata can serve as surrogates for the objects described.

2.3.2 Metadata schemas

In theory, the set of possible statements about any information object is infinite, and not all statements are useful in a given context. For this reason, metadata creation is typically guided by a set of rules. Metadata schemas contribute substantially to these standardization efforts, as they define what statements can be made about an information object, and how they can be made (Pomerantz, 2015). The terms *metadata schema* and *metadata standard* are often used synonymously. Although metadata schemas are not always standards acknowledged by official organizations like ISO, they are considered *de facto* standards if they are widely used (Hider, 2018).
Schemas cover different aspects of describing information objects, such as elements or values, including a set of elements (or fields), their meaning and relationships towards each other. Elements are the smallest structural components of metadata. Statements about information objects are made by assigning values to these elements. Metadata schemas may also apply constraints on the input of values, for example via controlled vocabularies or restrictions regarding data type, length or occurence. The *element set* comprises all elements defined by a schema. The basic unit in managing metadata is the *metadata record*, the set of all statements made about an information object (Zeng and Qin, 2016).

ID	DataCite-Property	Occ	Definition	Allowed values, examples, other constraints
1	Identifier	1	The Identifier is a unique string that identifies a resource. For software, determine whether the identifier is for a specific version of a piece of software, (per the Force11 Software Citation Principles[11]), or for all versions.	DOI (Digital Object Identifier) registered by a DataCite member. Format should be "10.1234/foo"

Figure 1: Excerpt from the DataCite Metadata Schema documentation, Version 4.3

Often, metadata schemas also clarify the terminology used (Hider, 2018). Some metadata schemas provide definitions or descriptions of each element set in their documentation (see, for example, the excerpt from the documentation of the DataCite Metadata Schema in figure 1 (DataCite Metadata Working Group, 2019; p. 13). As the meaning of terms may differ depending on disciplinary context, definitions can be very useful for creating and understanding metadata.

If new needs, practices or insights emerge, metadata schemas may have to be adapted. Therefore, many metadata schemas are regularly updated, or new schemas are developed (Zeng and Qin, 2016). Because metadata schemas restrict what statements can be made about information objects, they are often tailored to a specific community or use case (Hider, 2018). The need for a suitable standard in each use case can lead to a multitude of highly specialized standards. The Research Data Alliance Working Group Metadata Standards Directory compiled a list of metadata schemas used for research data that currently includes 98 standards.[1] (Ball et al., 2014)

Depending on the area of application, metadata schemas may be more or less specific. The specificity of a standard is the result of balancing various interests: should uniform statements be made about a large number of potentially diverse information objects, or should statements about more uniform information objects be particularly detailed? The DataCite Metadata Schema is intentionally generic to enable data retrieval and citation for a large variety of datasets (DataCite Metadata Working Group, 2019), whereas a discipline specific metadata schema like Darwin Core allows more detailed descriptions of datasets (Wieczorek et al., 2012). More general metadata schemas can be *qualified* to suit more specific information needs, for example by adding new (sub-)elements to the element set (Hider, 2018).

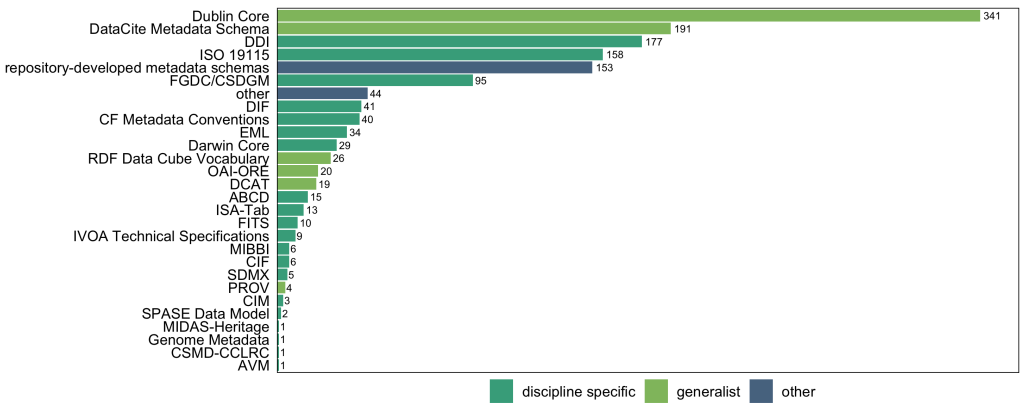

Figure 2: Metadata schemas adopted by repositories indexed in re3data

[1] source: RDA Metadata Standards Catalog API: https://rdamsc.bath.ac.uk/api/m (11.11.2020)

Considerations regarding the specificity of metadata schemas are also relevant for the selection of metadata schemas at research data repositories and other data service providers, as metadata schemas may significantly influence data discovery and interoperability. The FAIR Principles recommend describing datasets with rich metadata (principle F2), but the metadata schema should be widely used (principle I1) (Wilkinson et al., 2016). Repositories may adopt several metadata schemas to ensure detailed yet interoperable metadata. Among all repositories currently indexed in re3data, the generalist schema Dublin Core is most widely used, as figure 2 shows.[2]

To date, there is only one systematic and comprehensive analysis of metadata schemas for the description of research data (Willis, Greenberg, and White, 2012). The goal of the study was to analyze differences and similarities in design, objective, and scope of metadata schemas. Included in the analysis were documentations, related publications and implementation files of 9 schemas covering physical sciences, life sciences, and social sciences. The structural analysis uncovered considerable variance in the size of the element sets, comprising between 142 and 1802 elements. The results of the qualitative analysis showed that no objective was mentioned for all nine metadata schemas, however ten goals were mentioned by more than half of the schemas: schema extensibility (8), data interchange (8), data documentation (7), data retrieval (7), data publication (6), data archiving (6), comprehensiveness (6), schema flexibility (6), abstraction (5) and intra-scheme modularity (5). These objectives reflect the needs of the community it serves (for example, the need for expanding and adapting the schema), but also repositories and other infrastructures (for example, the need to exchange (meta-)data). An interesting point to note here is that the objective of comprehensive metadata descriptions was mentioned more often (6) than providing a sufficient or minimal set of schema elements (4), an indication that completeness of descriptions may be valued higher than their homogeneity. Although this study is the most comprehensive analysis of metadata schemas for research data, the sample still is small, and not representative of all metadata schemas applicable to the description of research data available today.

2.3.3 Metadata types

Depending on the function and characteristics of the statement being made about an information object, metadata elements can be grouped into types, for example according to the typology proposed by the United States National Information Standards Organization (NISO) shown in table 1 (Riley and National Information Standards Organization (U.S.), 2017; p. 6). In this typology, descriptive

Descriptive metadata	For finding or understanding a resource
Administrative metadata	
Technical metadata	For decoding and rendering files
Preservation metadata	Long-term management of files
Rights metadata	Intellectual property rights attached to content
Structural metadata	Relationships of parts of resources to one another
Markup languages	Integrates metadata and flags for other structural or semantic features within content

Table 1: Typology of metadata elements (Riley and NISO)

metadata capture essential information about an information object. In the case of research data, this would include the title, description or persistent identifier of a dataset. These statements enable the identification, discovery and retrieval of the dataset, and provide the necessary information to understand what the dataset is about. The NISO typology subsumes all information related to the

[2]sources: re3data API, RDA Metadata Standards Catalog API (11.11.2020)

management of an information object under the umbrella *administrative metadata*. This includes technical, preservation and rights metadata. Technical metadata cover information necessary to make use of digital files, for example the file format or media type of a dataset. Preservation metadata capture information related to the long-term management and availability of information objects, for example checksums to verify the integrity of a dataset. Rights metadata include statements about legal aspects of the information object, including the license. Relationships between information objects or parts of it are subsumed under structural metadata. Structural metadata include references to journal articles related to a dataset, or references from a collection to its parts. The NISO typology also includes a special case of metadata, markup languages. Markup languages enable the integration of metadata directly within the content of an information object.
Metadata typologies are useful, because schema elements can be matched to the categories described. This provides an overview of objectives and functions metadata schemas cover, and what types of statements a schema emphasizes. For example, Figure 3 shows that the elements of the DataCite Metadata Schema (Version 4.3) cover all metadata types except for preservation metadata and markup languages, and that there is a strong focus on descriptive metadata elements.

Figure 3: Elements of the DataCite Metadata Schema, Version 4.3, grouped by the NISO metadata typology

A detailed list of the elements in the DataCite Metadata Schema and the corresponding NISO metadata types was compiled and can be found in the Appendix A.
Missing in the NISO typology, but highly relevant in the context of research data, are metadata functions related to good scientific practice. In documenting decisions made throughout the research process, metadata support verification of results (Leonelli, 2016). If thorough documentation is not provided, researchers cannot reliably reuse data, even in critical situations like pandemics (Schriml et al., 2020).

2.3.4 Characteristics of metadata

Metadata are not uniform, but have different characteristics depending on how they are created (Gilliland, 2008).
Metadata can be generated automatically or manually. Currently, there are only few examples of automated metadata creation for research data besides adding technical or intrinsic metadata. However, there are several conceivable areas of application as automating approaches mature (Weber et al., 2019).
The degree of structure metadata provide may also vary. Although standardized metadata are a prerequisite for interoperability and comprehensive discovery services, less structured descriptions can convey necessary context for understanding and reusing data. One example of unstructured metadata are data papers, which mirror traditional scientific publication formats, but focus on an in-depth description of data collection and processing (Candela et al., 2015).

Information objects can also be described at different levels, for example at the collection level, at the level of individual datasets, or even more fine-grained at the level of individual variables. The granularity of descriptions is a relevant topic in the context of research data management, as research has shown that one dataset (a packaged item in a repository) often comprises several files (Wehrle and Rechert, 2018; Bugaje and Chowdhury, 2017).
Characteristics also include the status of metadata, particularly whether descriptions are regarded as static, unchangeable objects (*metadata-as-product*), or as a dynamic and iterative process (*metdatata-as-process*) (Edwards et al., 2011). Conceptualizing metadata as a process is more suitable to the social world of research, where metadata descriptions may be generated incidentally at first, and refined later as scientific communication evolves (Edwards et al., 2011). Information infrastructures tend to view metadata as immutable products. However, an ethnographic study of the practices at a research data repository showed that metadata records are created in cooperation and improved upon in iterative processes (Plantin, 2018). Metadata management approaches treating metadata records as unchangeable objects may limit their usefulness across the entire data life cycle and to different data users (Habermann, 2018a).

2.3.5 Metadata and research data

Key characteristics of the information object described should be reflected in metadata, as metadata are often used as a surrogate for the information object they represent. Were these characteristics not sufficiently documented, information objects could be misrepresented, overlooked or misunderstood. Gilliland identifies three features of any information object that are relevant for metadata descriptions – content, context, and structure:

> "*Content* relates to what the object contains or is about and is *intrinsic* to an information object. *Context* indicates the who, what, why, where, and how aspects associated with the object's creation and subsequent life and is *extrinsic* to an information object. *Structure* relates to the formal set of associations within or among individual information objects and can be *intrinsic*, *extrinsic*, or *both*." (Gilliland, 2008)

Taking data as an example, it becomes clear that not only content but also context and structure should be considered in the description of information objects. Datasets are often conceptualized as self-contained items that can be copied, transferred, understood and reused by anyone. However, this notion must be challenged, as the discussion of the representational view of data in this thesis demonstrates. An alternative to the self-contained *dataset* is to consider the "data setting" - the social and technical context of a dataset (Loukissas, 2019). Metadata convey these data settings, for example the circumstances under which data were generated, used as evidence, managed, or shared. Throughout their lifetime, data are often moved across contexts, or *settings*, to be reused. Sabina Leonelli calls these movements *data journeys* (Leonelli, 2020a). According to Leonelli, research data can travel because they are *mutable*, meaning that, to a certain extent, they can be transformed to fit a specific use, and this mutability depends in large parts on the metadata available.
Infrastructures enable data journeys by reducing friction between individual practices and the community: "An infrastructure occurs when the tension between local and global is resolved." (Star and Ruhleder, 1996; p. 114)
Discussions surrounding the elements a description should contain are ongoing. For example, the digital preservation community introduced the concept of *significant properties*, the set of properties that must be preserved to ensure ongoing access to and usability of a digital object, and the information necessary to transport its meaning (Recker and Müller, 2015). Particularly with regard to enabling

future data reuse, this concept also has applications in research data management. However, what characteristics of data are considered *significant* differs across disciplines, and the properties required for successful data reuse go beyond what is necessary for digital preservation (Faniel and Yakel, 2011).

2.3.6 Metadata quality and metadata evaluation

As described above, metadata sustain most functions of information infrastructures, and it is often via metadata that users first come in contact with information objects. Therefore, metadata quality is an important factor for operating information infrastructures and ensuring a good user experience. The International Organization for Standardization (ISO) defines quality as the "degree to which a set of inherent characteristics of an object fulfills requirements" (International Organization for Standardization, 2015). Which requirements must be met depends on the nature of the object and the expectations attached to it. In the case of metadata quality, several requirements can be applied. For example, Bruce and Hillmann name the following (the descriptions are paraphrased based on Bruce and Hillmann, 2004; p. 5ff):

- **Completeness** The information object should be described as completely as possible, and as many elements as possible in the element set should be used.
- **Accuracy** The information object should be described accurately. The statements should be factually correct and conform to the same rules, for example adhering to a uniform structure for writing names.
- **Provenance** The information object should be described to include information on data collection and metadata creation, as well as transformations and other changes.
- **Conformance to expectations** The information object should be described, to the extent possible, to meet the expectations of potential users, for example by including all information potential users may expect, without any irrelevant statements.
- **Logical consistency and coherence** The information object should be described by elements that are consistent with standard definitions, and descriptions should be coherent across collections.
- **Timeliness** The information object should be described in its current form, and descriptions should be disseminated with as little delay as possible.
- **Accessibility** The information object is described by metadata that can be easily accessed, without any physical or intellectual barriers.

When evaluating metadata quality, the conformity to a set of requirements is determined. The unit of reference for metadata evaluation can either be individual metadata elements, metadata records or entire metadata collections (Zeng and Qin, 2016). Metadata quality evaluation requires the operationalization of these requirements, for example in the form of metrics that describe what is measured and how (Palavitsinis, 2013). When considering the requirements proposed by Bruce and Hillmann, it becomes clear that no one approach or metric would be sufficient to evaluate all of them. For example, metadata records alone may suffice to evaluate certain quality dimensions, like *completeness*, whereas additional information from potential users would be required for *conformance to expectations*.

2.3.7 Metadata evaluation in the context of research data

Metadata quality for research data has been a research topic for several years (Rousidis et al., 2014). However, the body of literature on metadata evaluation for research data specifically is still limited. While there are ambitious policies in place and the number of repositories is growing, little is known

about the status quo and results of data stewardship (York, Gutmann, and Berman, 2018). An analysis of data deposit requirements of 20 repositories catering to archaeology, zoology and quantitative social science revealed significant differences in metadata requirements within and across the disciplines (Kim, Yakel, and Faniel, 2019). There is also tentative evidence for differences in metadata practices at institutional repositories (Manninen, 2018). However, this study was based on only one metadata record each from 15 institutional research data repositories using the repository platform Digital Commons. Overall, little is currently known about differences in metadata practices across institutions, limiting data aggregation and comparative analysis (Gregg et al., 2019).

A lot of the research available in the area of metadata evaluation is driven by the goal to establish metrics, for example metrics measuring the impact of datasets (Cousijn et al., 2019; Robinson-Garcia et al., 2017; Peters et al., 2016). Metrics require comparability and therefore pose particular challenges to metadata records, such as the granularity of descriptions or the documentation of versioning (CODATA-ICSTI Task Group on Data Citation Standards and PractOut of Cite, 2013).

Data discovery is another field advancing research on metadata collections describing research data. Currently, data discovery is based on structured metadata only (Chapman et al., 2019). Therefore, the quality of metadata is of particular importance in the context of data discovery. Discovery can also be a driver for metadata quality, as metadata aggregation from various sources requires uniform and high-quality metadata (Bruce and Hillmann, 2004). Data discovery gained considerable attention with the launch of Google Data Search in 2018 (Noy, 2018). Google Data Search is neither the only nor the first discovery service for datasets (Chapman et al., 2019), but it has the potential to make data discovery and reuse more accessible to a wider audience. The service uses standards for embedding metadata in web pages (schema.org and DCAT) and crawls for web pages providing metadata for datasets (the schema.org class *Dataset* or comparable DCAT concepts)(*Google Developers: Dataset*). In August of 2020, the Google Dataset Search index included more than 31 million datasets (Noy, 2020). The schema.org class *Dataset* only has two required properties: *title* and *description*. These two properties are present in all indexed metadata records, whereas other properties are used much less frequently: for example, license information is available for 34.80 % of all datasets, and only 11 % are assigned a DOI (Benjelloun, Chen, and Noy, 2020). The services acknowledges that metadata quality presents a major challenge and is exploring approaches towards resolving this issue (Benjelloun, Chen, and Noy, 2020).

Currently, the DataCite Metadata Store is one of the most comprehensive sources for metadata on research data, and its use is not restricted. In 2017, Robinson-Garcia et al. evaluated all metadata records in the DataCite Metadata Store at that time (7440415 records) with regard to the completeness of individual metadata records and the level of standardization provided (Robinson-Garcia et al., 2017). The authors found that most metadata records were not complete, for example only 51 % of metadata records provided information on the language of the resource. Least common was information on relations to other resources (25 %) and contributors (18 %). In some cases, even mandatory fields were empty. The authors also analyzed the content of metadata properties to evaluate the level of standardization achieved. They found that there was some overlap of schema elements (for example between *publicationYear* and *date*), potentially reducing clarity and ease of analysis of the descriptions, and free text fields led to less uniform entries.

Research data repositories are important actors in metadata and management. They are not just containers for datasets, but take on an active role in institutionalizing research data management (Mayernik, 2015). Several studies have investigated repositories' practices and their potential influence on metadata for research data. An analysis of metadata elements related to aspects of data sharing (for example availability, coverage, format etc.) offered by 5 generalist repositories revealed heterogeneity in the number of supported metadata elements, the obligation levels (mandatory or not), as well as the use of controlled vocabularies (Assante et al., 2016).

Recently, Löffler et al. conducted a detailed and systematic assessment of metadata collections at five research data repositories: three generalist repositories (DRYAD, Figshare, Zenodo) and two discipline specific repositories catering to datasets relevant for biodiversity research (PANGAEA, GBIF) (Löffler et al., 2020). They calculated the ratio of metadata records including information on categories that were identified as relevant to biodiversity research (for example environment, organism, location or data collection method). According to the study, generalist repositories were more likely to use basic metadata schemas such as DublinCore and the DataCite Metadata Schema. The authors found that in the case of DRYAD, the majority (9 out of 15) of metadata properties of DublinCore were used in at least 80% of the metadata records, other elements were used less frequently. Similar patterns were observed for the other generalist repositories as well. In Zenodo, only 45 % of records provided subject information. Although the two discipline specific repositories in the sample used more domain-relevant schemas, metadata records were more often incomplete. However, the assessment was based on different metadata schemas, therefore results can not be directly compared across repositories.
Metadata quality varies within and across two discipline specific research data repositories storing data about samples used in biomedical experiments (Gonçalves and Musen, 2019). Representations of attributes, for example geographic locations and time, were found to be heterogenous even within one repository, potentially limiting dataset findability.
Quarati and Raffaghelli analyzed the quality of metadata records in the generalist research data repository Figshare, and found no correlation between metadata quality and views or downloads of datasets (Quarati and Raffaghelli, 2020). Their findings suggest that other factors besides metadata quality influence data reuse, but this question needs further investigation. For example, infrastructural, regulatory, and socio-technical factors impeding data reuse could be considered in future research (Bates, 2018).
In a blog post for the DataCite Blog, Ted Haberman applied metrics for describing metadata to the collections of more than 1200 data centers in the DataCite Metadata Store (Habermann, 2018b). He finds that most data centers use more than the mandatory schema elements required by the DataCite Metadata Schema, but metadata collections are far from being complete or homogenous. His findings hint at a loose negative correlation between metadata completeness and homogeneity at the repository level, meaning that descriptions tend to get more heterogenous the more schema elements are being used, but the relationship between the two metrics is not further quantified.
The studies discussed here contribute to the understanding of metadata for research data. However, there is a lack of systematic analyses that take into account aspects such as institutional differences or repository characteristics.

2.3.8 Information behavior research in the context of research data

In recent years, several studies applied models and theories of information behavior research to research data. This approach could inform future evaluations of metadata for research data. If repositories understand how data users search for, assess and interact with datasets, they can better respond to data users' information needs. For example, metadata schemas could be adapted to better reflect researchers' information needs (Löffler et al., 2020). Data users' information needs and behaviors have been studied in the context of data reuse (see for example Faniel, Frank, and Yakel, 2019). In recent years, other aspects of information behavior have been analyzed, including searching for and making sense of datasets.
Metadata quality has a large impact on the findability of data, and research data repositories should consider their users' information needs and diverse search strategies when describing datasets (Gregory et al., 2020b). However, metadata are not the only source of information on datasets available

to researchers. Researchers also rely on supplemental materials, text publications, and personal connections when searching for and making sense of data (Gregory et al., 2019). Data users' information needs may also differ depending on the type of intended use, and in what stage of the research process data are used (Gregory et al., 2020a).
Based on interviews with 20 data workers, Laura Koesten and her co-authors developed a framework for human interaction with structured data (Koesten et al., 2017). Among other findings, they identified information needs relevant to evaluating whether a dataset can be used. Table 2 summarizes information needs associated with assessing the relevance, usability, and quality of datasets (Koesten et al., 2017; p. 1283). Ideally, these information needs should be reflected in metadata schemas and captured by in metadata records.

Assess	**Information needed about**
relevance	context, coverage, original purpose, granularity, summary, time frame
usability	labeling, documentation, license, access, machine readability, language used, format, schema, ability to share
quality	collection methods, provenance, consistency of formatting / labeling, completeness, what has been excluded

Table 2: Information needs for selecting datasets

In a later study, the authors build on these results to study data-centric sensemaking activities (Koesten et al., 2020a). They identify a large variety of information structures supporting sensemaking activities, as well as information needs that should be reflected in these information structures from the perspective of researchers *close* or *far* from datasets. The working group also studied data summaries. Common attributes in 150 data summaries that were in part crowdsourced and 269 data search diaries were compared (Koesten et al., 2020b). The authors found significant overlap between the attributes of summaries and search diaries, and propose a framework for crowdsourcing or automating data summarisation. The most common attributes could also inform the development of metadata schemas. The same group of researchers conducted logfile analysis of data portals. Results show that temporal and geospatial coverage of datasets may be of particular relevance to data search (Kacprzak et al., 2019). Repositories can benefit from this research by adapting their services to better fit users' information needs. Data discoverability could be improved, for example, by providing multiple search interfaces, increasing interoperability with other services, and offering metadata for harvesting (Wu et al., 2019).
Information behavior research offers a promising perspective on metadata for research data. However, reports on the applicability of these findings to repository practices are currently missing.

2.3.9 Metadata practices

Data practices are "[...] the work involved in creating, managing, and using research data and their associated metadata." (Mayernik, 2015; p. 1) Data production, the process by which usable data is created and packaged for submission to a repository, also includes the creation and management of metadata (Baker and Mayernik, 2020). Those actions directly related to the creation, maintenance and dissemination of metadata can be referred to as metadata practices (Mayernik, 2015). Similar to variations in data practices across disciplines (Leonelli, 2020a), metadata practices vary across

and within organizations or disciplines (Mayernik, 2015). The extent and characteristics of these differences remain, however, largely unknown.
Metadata creation in practice often is incremental, and the process may be distributed across several organizational units (Baca, 2016). As a result of changing circumstances and information needs, metadata records often need to be revised and modified; metadata labor, however, frequently remains invisible to outsiders (Downey, Eschenfelder, and Shankar, 2019). According to a group of authors commenting on metadata practices related to research on the ongoing COVID-19 pandemic, metadata labor is often taken for granted or undervalued, and good metadata practices should be encouraged by various stakeholders (Schriml et al., 2020). The authors conclude with highlighting the value of structured metadata: "[it] is an unglamorous corner of science, but metadata standards are vital infrastructure – often holding the key for data-driven research discoveries." (Schriml et al., 2020)

2.4 Situating the research questions

As the discussion of existing definitions showed, *research data* can be conceptualized as a relational category that is applied to information objects. In this understanding, the evidential value of data depends on their application. The documentation of research data in the form of metadata is of particular importance since metadata capture the context of data and facilitate data (re-)use. *Research data repositories* are information infrastructures specialized on data. Like other information infrastructures, they create and manage *metadata* that support most repository functions.
Existing research offers limited information about the relationship between repositories and metadata for research data. Therefore, following a quantitative assessment of metadata quality, this thesis aims at making distinctive features of metadata for research data visible (RQ1). Metadata collections of individual research data repositories and the potential influence of repository characteristics on these collections are investigated (RQ2). To gain an insight into processes related to metadata creation and revision, aspects of changes to metadata records will also be considered in the analysis (RQ3).

3 Methodology

The thesis is based on infrastructural inversion, a concept from the infrastructure studies. Infrastructural inversion as described by Bowker and Star is a means to make infrastructures visible that otherwise often fade into the background (Bowker and Star, 2000). Focusing on one layer of infrastructures for research data, this thesis attempts to disentangle the interrelation between research data repositories and structured metadata (here: metadata adhering to the DataCite Metadata Schema). The aim is to gain a better understanding of the metadata products of and, to a lesser extent, metadata practices within research data repositories.
Infrastructure studies can use a variety of methods, depending on the phenomena under investigation and the research questions (Bowker and Star, 2000). Because a systematic approach to the topic is currently lacking, this thesis follows a quantitative approach that is based on the joint analysis of two sources providing comprehensive information on research data repositories (re3data) and metadata for research data (DataCite) respectively.
The following sections describe the steps preceding and including the analysis, specifically matching procedures between the two data sources, selection criteria for repositories, metadata harvesting, and data processing. Indicators used to describe metadata collections are outlined.
Metadata records were collected and analyzed using the statistical software R.

3.1 Data source selection

This thesis is based on data provided by two services: DataCite and re3data. These data sources were selected because of the broad coverage of repositories and datasets they provide, because the data is made available via open interfaces and because there is already some degree of interoperability between the two services – see for example the reference to re3data IDs in the DataCite client information described below. Other data providers, such as bibliographic databases, did not fulfill these requirements to the same degree.

DataCite

DataCite is an international not-for-profit organization providing DOIs for research data and other research outputs (*DataCite's Value*). An organization may become a DataCite member and use DataCite services to register DOIs for their clients (*DataCite - Members*).
On this infrastructure for assigning persistent identifiers, DataCite built a number of services promoting the use of research data. For example, clients submit metadata adhering to the DataCite Metadata Schema to DataCite when registering a DOI for a dataset. DataCite aggregates these metadata records and offers them for harvesting via several interfaces, for example the DataCite REST API. An analysis from 2017, discussed in detail above, demonstrates that DataCite currently holds the most comprehensive collection of metadata records on research data (Robinson-Garcia et al., 2017).

re3data

re3data is a comprehensive registry for research data repositories, covering all disciplines and repository types (Kindling et al., 2017). Repositories are described by an editorial board based on the re3data Metadata Scheama (Rücknagel et al., 2015). Metadata records of all repositories are offered for harvesting via an open API.

3.2 Repository Selection

Repositories were selected based on general selection criteria as well as criteria specific to the matching procedures. The sample of repositories was compiled on August 3rd, 2020.
Selected repositories must be listed in re3data. Using the re3data API, the selection was restricted to repositories using a specific metadata schema to ensure comparability across all metadata records. For this thesis, the DataCite Metadata Schema was selected (*metadataStandards* = "DataCite Metadata Schema"). The schema is widely used by repositories, as well as the DataCite Metadata Store, the source for metadata records used in this thesis. Additionally, selected repositories must assign DOIs to datasets (*pidSystems* = "DOI"). Since DataCite is the organization assigning DOIs for datasets, repositories adopting this identifier system transfer their metadata to the DataCite Metadata Store upon DOI registration, ensuring that metadata records can be harvested.
In addition to these general selection criteria, repositories were selected based on the ability to match re3data repository descriptions and DataCite metadata records. These additional criteria guarantee the joint analysis of the two data sources. The two matching processes applied are described in the following section. Depending on the matching process used, some repositories were excluded from the sample based on additional criteria, particularly if no operational OAI-PMH interface unique to the repository could be found.
Repositories without published datasets were excluded from the analysis, as well as repositories only publishing text publications. As Robinson-Garcia et al. argued, text publications can reasonably be excluded from analysis based on data from the DataCite Metadata Store (Robinson-Garcia et al., 2017). Furthermore, academic social networks (for example researchgate) were removed from the sample, as they do not fall within the scope of research data repositories as they are understood in this thesis.
Details on the sampling process for each matching procedure, along with a summary of sample characteristics, are described below.

3.3 Data Collection

3.3.1 Matching of re3data repository descriptions and DataCite metadata records

Metadata records in DataCite and repository descriptions in re3data are currently not directly linked. In part, this is due to the lack of a persistent identifier for research data repositories that is used in both sources. The field *publisher* of the DataCite Metadata Schema is a free-text field, therefore entries are very heterogenous (Robinson-Garcia et al., 2017) and matching with re3data records is complicated. Design differences of the two databases also contribute to this problem: although both re3data and DataCite provide means of uniquely identifying entities hosting research data, they may

refer to different types of entities. Whereas re3data IDs refer to individual research data repositories, DataCite client IDs often refer to institutions that can host more than one repository. Therefore, the DataCite client IDs are not necessarily unique to one repository.
Due to these design differences, a workaround is required to link the two data sources. By establishing a reliable link between re3data and DataCite, metadata records can be retrieved from the DataCite Metadata Store based on attributes of repositories in re3data.
Existing research describes two approaches to establishing this link:

matching by DataCite client information

One approach to matching re3data repository descriptions and DataCite metadata records is based on DataCite client information (Habermann, 2018b). DataCite clients, for example research data repositories, submit metadata records to DataCite in order to register DOIs for datasets. Information on DataCite clients can be accessed via the DataCite client API.[1] As described above, DataCite clients do not always correspond to one specific repository. Whenever that is the case, a DOI resolving to the re3data entry of the repository was added to the DataCite client information. Therefore, these *re3data DOIs* can be resolved to establish a link between DataCite clients and re3data IDs. The client IDs correspond to sets defined in the OAI-PMH interface of the DataCite Metadata Store, and metadata collections of repositories can be harvested by client ID.
For this thesis, information for all DataCite clients was first retrieved via the DataCite client API. The results were then filtered based on the presence of the element *re3data*, which contains the re3data DOIs for DataCite clients directly corresponding to a repository listed in re3data. These re3data DOIs were then resolved in order to obtain the URLs of the re3data records of the repositories. The re3data ID, which is part of each re3data URL, was then extracted. The result was a lookup table containing all DataCite client IDs that have corresponding re3data IDs listed in the DataCite client information. To apply the repository selection criteria to this lookup table, a list of all re3data IDs meeting the general sampling conditions (metadata schema and persistent identifier system) was retrieved from the re3data API. This list was then intersected with the lookup table, resulting in a list of all repositories meeting the general selection criteria that also have a DataCite client ID. Figure 4 illustrates the matching process described.

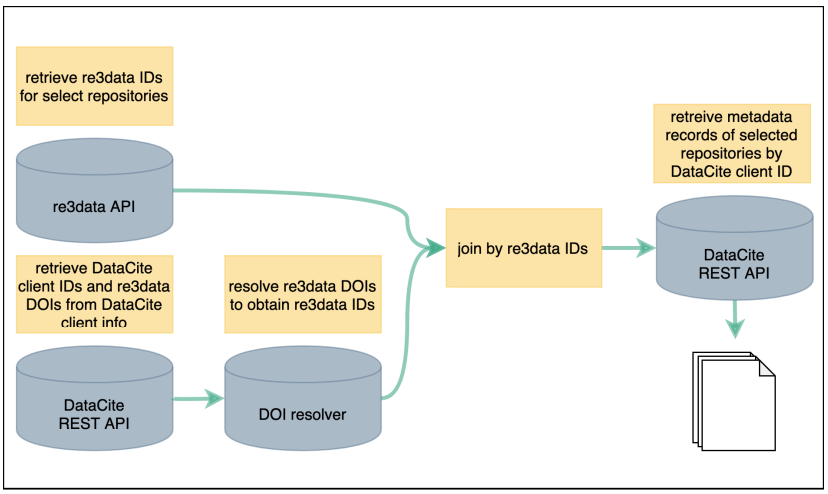

Figure 4: Matching process by DataCite client information

[1]The DataCite client API is available at: https://api.datacite.org/clients

Since this method only relies on DataCite, a central and reliable infrastructure, it was treated as the preferred method in case both matching processes could be applied to one repository.

matching by harvesting DOIs from repository APIs

Another approach is based on harvesting DOIs of all datasets via the selected repositories' APIs, for example OAI-PMH interfaces, and then retrieving metadata records for these DOIs from the DataCite Metadata Store (Löffler et al., 2020; Weber and Kranzlmuller, 2018). In this case, the link between re3data and DataCite is established intermediately by harvesting DOIs via repository APIs. Figure 5 illustrates the matching process described.

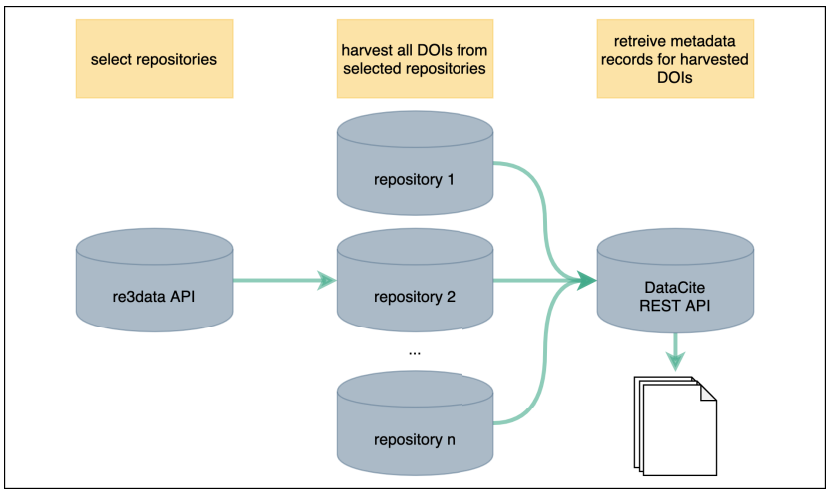

Figure 5: Matching process by harvesting DOIs from repository APIs

This approach is limited by the fact that not all repositories in re3data offer their metadata for harvesting, and some of the links to OAI-PMH interfaces mentioned in re3data are no longer operational. Therefore, matching via DataCite client information was preferred if possible.
First, the re3data API was queried in order to retrieve repositories matching the general sampling conditions (metadata schema and persistent identifier system) that also provide an OAI-PMH interface (*apis* = "OAI-PMH"). The resulting list was intersected with the repositories successfully matched via DataCite client information, which was treated as the preferred method. For the remaining repositories, the API URL(s) were retrieved via the field *api* for manual checking. If a link was broken or the interface no longer operational, the repository website was searched for alternative endpoints. If no operational OAI-PMH interface could be identified, the repository was excluded from the analysis. In some cases, URLs to endpoints were duplicated in the sample. These repositories all appeared to rely on figshare as their technical infrastructure, and neither the OAI-PMH interface nor sets could be uniquely attributed to one repository. Therefore, these repositories were excluded from the sample as well.
Metadata records of all of the operational OAI-PMH interfaces were harvested. DOIs were then extracted from the retrieved XML records and saved for harvesting. For some DOIs, the DataCite interfaces did not provide any information. Likely, these DOIs were registered with another organization such as Crossref, and therefore excluded from this analysis.

3.3.2 Harvesting metadata records from the DataCite Metadata Store

After both matching processes were concluded, metadata records were harvested via the DataCite OAI-PMH interface between August 3rd and August 10th, 2020. Depending on the matching process, different OAI-PMH verbs were used:

- **matching by DataCite client information:** the OAI-PMH verb *ListRecords* was used to extract sets of metadata records. The OAI-PMH sets of the interface correspond to DataCite clients, therefore the DataCite client ID can be used to obtain all metadata records of specific clients. Using this method, all metadata records of repositories meeting the selection criteria were harvested using the DataCite client IDs.
- **matching by harvesting DOIs from repository APIs:** the OAI-PMH verb *GetRecords* was used to extract individual metadata records via the DOIs obtained in the matching process.

In both cases, an upper time limit was defined, limiting results to all metadata records registered up to and including July 31st, 2020 ("2020-07-31T23:59:59Z").

3.3.3 Information on changes to metadata records

DataCite tracks the history of metadata records registered since March 10th, 2019 and offers this information via the *activities* endpoint.[2] In offering this administrative information, DataCite makes changes to metadata records transparent and provides insights into metadata practices related to the description of research data. In this thesis, the information offered by the activities endpoint is analyzed with a focus on RQ3. Specifically, the version number of each metadata record is used, if available, to investigate whether repository characteristics have an influence on the number of changes to metadata records.
For all metadata records registered after March 10th, 2019, the version number was harvested via the DataCite activities endpoint. Metadata records registered before that date were excluded to limit the analysis to metadata records with a complete provenance record. The activities endpoint was queried on August 7th, 2020. A version number was available for 164775 datasets from 32 repositories.

3.4 Data Processing

3.4.1 Processing DataCite metadata records

Using the method described above, raw XML files were harvested from the DataCite OAI-PMH interface. Information was then extracted using XPath expressions. The XPath expressions were formulated based on the schema documentation as well as xsd and xml examples provided by DataCite. For all schema elements mentioned in the current schema version, the occurence was extracted. Additional variables were extracted from the metadata records to expand the scope of the analysis. For the schema elements *title* and *description*, the number of characters were included in the analysis. Because a metadata record can contain more than one title and description, the number of characters were summed up for all instances of these elements per metadata record. The text of the element *publicationYear* was extracted for a time-based analysis of metadata records. To answer RQ3, the

[2]Information on the DataCite activities endpoint is available at: https://support.datacite.org/docs/tracking-provenance

number of changes to metadata records registered with DataCite was included in the analysis.
Additional information was collected from the DataCite OAI-PMH interface for the purposes of data cleaning and preparation for the analysis.
Since the publication of the first version, the DataCite Metadata Schema has been adapted several times to add new elements. These revisions result in varying sizes of element sets across schema versions. Therefore, the stated schema version was extracted for all metadata records to test whether this information could be used to refine the analysis. Knowledge of the schema version of metadata records makes their assessment more precise, since the total number of elements available for descriptions can be determined more accurately for each metadata record. However, information on the schema version was not available for many metadata records, and if present, very heterogenous. For example, many records stated implausible numerical values. Information on the schema version could therefore not be used to determine the number of metadata elements in the element set more accurately. As a result, the *datestamp* provided by DataCite was used to approximate the schema version. The datestamp specifies the date when a metadata record was first registered with DataCite. The assumption underlying this approximation of the schema version is that a metadata record always follows the latest available version of the schema. Release dates of the schema versions were retrieved from the DataCite website (*DataCite Metadata Schema*). The release dates and sizes of the element sets for each schema version are listed in Table 3.

schema version	release date	size of the element set
4.3	2019-08-16	83
4.2	2019-03-20	76
4.1	2017-10-23	72
4.0	2016-09-19	66
3.1	2014-10-16	44
3.0	2013-07-24	42
2.2	2011-07-01	31
2.1	2011-03-28	31
2.0	2011-01-24	31

Table 3: Version history of the DataCite Metadata Schema

The value of the schema element *resourceTypeGeneral* was used to exclude text publicatons from the analysis. According to version 4.3 of the schema, the resource type *text* is described as "A resource consisting primarily of words for reading", for example: "Grey literature, lab notes, accompanying materials, data management plan, conference poster." (DataCite Metadata Working Group, 2019; p. 40) This indicates that resources of the type text do not fall under the definition of research data used in this thesis. The analysis of Robinson-Garcia et al. supports this assumption. They found that most resources with the *resourceTypeGeneral* text were manuscripts, conference papers and journal articles, and recommend filtering DataCite metadata records based on their resource type before conducting analysis (Robinson-Garcia et al., 2017). In order to reliably exclude text publications from the analysis, metadata records missing information on the resource type were removed.
In their analysis of the DataCite Metadata Store, Robinson-Garcia et al. also found many empty metadata records, which are likely a result of internal metadata management processes at research data repositories (Robinson-Garcia et al., 2017). Therefore, the harvested xml files were checked for empty metadata records, but none were found.

3.4.2 Adding repository information from re3data

Metadata for all repositories included in the analysis was retrieved via the re3data API. Following an exploratory approach, several metadata elements were retrieved for an analysis of the impact of repository characteristics on characteristics of metadata collections. The issues and challenges related to research data repositories described in the literature section of this thesis guided the selection process. Repository characteristics extracted include repository types, subjects, and certification.
The certification status of repositories was manually verified. At the moment, several initiatives are fostering repository certification[3], therefore the certification status of repositories may change quickly. The information extracted from re3data was accurate and could be verified.

3.5 Sampling process and sample characteristics

The repository selection process was outlined above. This section provides specific information of the sampling process and a detailed description of the resulting sample of repositories. The repository sample was compiled on August 3rd, 2020.

3.5.1 Matching by DataCite client information

On August 3rd, 2020, 202 of the 2169 clients listed in the DataCite client API had listed a re3data DOI, and 177 repositories in re3data met the general selection criteria described above. The intersection of both sets comprises 49 repositories.
After checking the DataCite API, 3 empty repositories were removed from the list. Additionally, one academic social network (researchgate) was excluded. After harvesting, text publications (*resourceTypeGeneral* = "text") were removed in preparation for the analysis. 4 repositories were removed from the list because they only contained text publications.
The remaining 41 repositories were added to the sample.

3.5.2 Matching by harvesting DOIs from repository APIs

As of August 3rd, 2020, 48 repositories in re3data met the general selection criteria described above and provided an OAI-PMH interface. After removing duplicates (repositories where both matching mechanisms could be applied), 26 repositories remained in the list. For 10 repositories, no operational OAI-PMH interface could be found, and in 4 cases, the OAI-PMH interface was not uniquely attributable to one repository (these repositories referred to the general figshare API). These repositories were excluded from the list. Metadata for the remaining 12 repositories was harvested via the OAI-PMH interfaces. The DOIs of all documents were extracted and cleaned if necessary. 7 repositories did not offer DOIs for harvesting and were therefore excluded from the list. Of the remaining 7 repositories, 2 also used the PID system handle, but only DOIs were retrieved. 1 repository also registered content with Crossref, and only DataCite DOIs were included. After harvesting, text publications (*resourceTypeGeneral* = "text") were removed in preparation for the analysis. 1 repository was removed from the list because it only contained text publications (this was the repository also registering content with Crossref). The remaining 6 repositories were added to the sample.

[3]For example, the EU project FAIRsFAIR supports repositories seeking certification: https://www.fairsfair.eu/fair-certification

3.5.3 Repository sample characteristics

In total, the repository sample comprises 47 repositories. Figure 6 shows the geographical distribution of institutions affiliated with the repositories in the sample (r3d:institutionCountry). A total of 56 institutions from 18 countries or regions are affiliated with the selected repositories. Most institutions are based in Europe (31) and North America (10). 3 institutions are based in Asia, 2 in Oceania, 1 in Africa, and none in South America. Countries of institutions affiliated with repositories in the sample are highlighted in the map (shown in dark green). One repository is only affiliated with an institution of the European Union (shown in light green). 9 international institutions are affiliated with repositories, 1 repository is only affiliated with an international institution (not shown on the map).

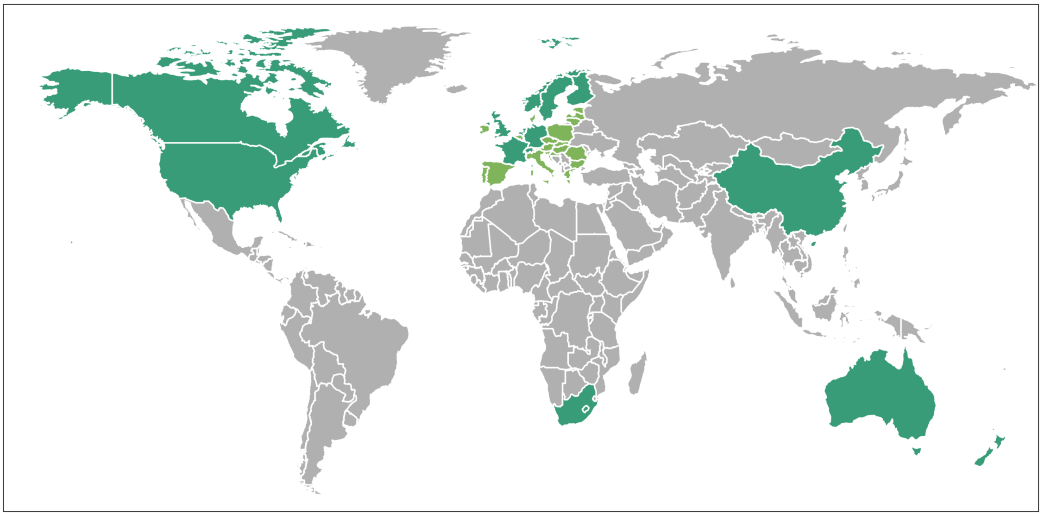

Figure 6: Countries of institutions affiliated with selected repositories

Figure 7 shows the subject groups of the repositories in the sample. The subject groups shown are derived the subject field in re3data (*r3d:subject*), which is based on the subject classification of the Deutsche Forschungsgemeinschaft (DFG).[4] In order to harmonize subject affiliation for all repositories, the notations of the DFG subject classification were shortened to one integer (the broadest category in the classification). Duplicate mentions of subject groups were removed for all repositories. The Venn diagram shows intersections between the subject groups for all repositories in the sample. For 27 repositories, notations from all four subject groups are mentioned in re3data. Other combinations of subject groups are much less common. 5 repositories only cover natural sciences, 3 humanities and social sciences, and 2 life sciences. There is no repository focusing only on engineering sciences in the sample.

[4]The DFG subject classification is available at: https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp

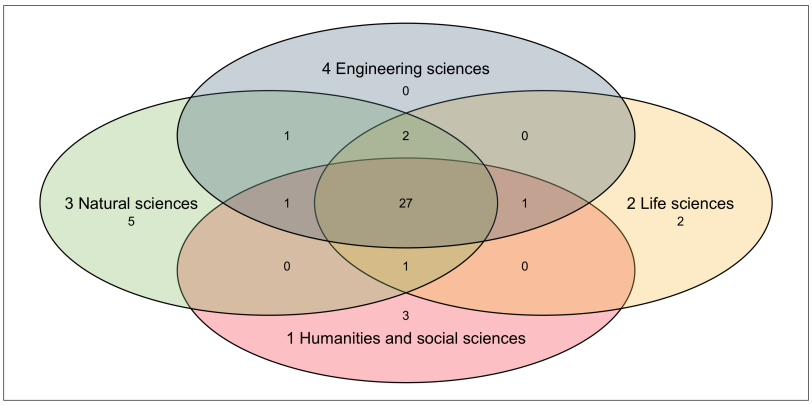

Figure 7: Subject group (combinations) of repositories in the sample

Figure 8 illustrates the types of repositories represented in a Venn diagram.

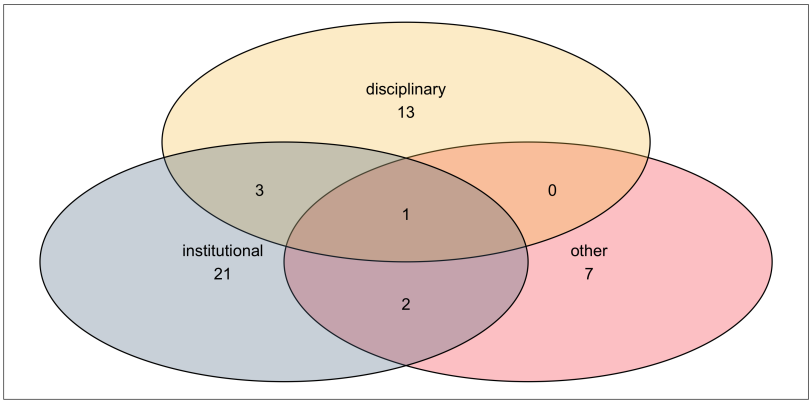

Figure 8: Type (combinations) of repositories in the sample

Most repositories in the sample are institutional (21) or disciplinary (13), and some are specified as *other* (7). 6 repositories do not fall in one category, but state a combination of the three categories, in one case all three categories.
Due to the matching process described above, the sampling strategy of this thesis is highly dependent on the technical maturity of repositories (DataCite client status or OAI-PMH interface). As a result, the sample is not representative of all repositories, as is illustrated by the geographical distribution that is strongly skewed towards Europe. In addition, the selected repositories demonstrate a surprisingly low degree of subject specificity, as most repositories cater to all four subject groups.
A full list of repositories in the sample can be found in the Appendix B.

3.5.4 Metadata sample characteristics

The metadata sample comprises a total of 606091 metadata records. 600047 metadata records in the sample are unique. This discrepancy is not caused by duplicates within the collection of one repository, but across the collections of two repositories. 6044 DOIs were present in the metadata collections of two repositories (in total, four repositories did contain duplicates). Further investigation showed that most intersections (6039 DOIs) occurred between two repositories, the remaining DOIs

(5) were duplicated across the two other repositories. The duplication of metadata records may be a result of (meta-)data reuse across repositories, or an institution may maintain several repositories and make datasets accessible via both infrastructures. In any case, collection management at research data repositories is intentional, and the duplicated metadata records are considered a valid part of both repositories' metadata collections. Therefore, duplicated metadata records were not removed. Figure 9 shows that repositories contain between 11 and 170201 metadata records, with a median of 561 and an average of 12895 records. As the smoothed density plot indicates, the metadata sample is skewed towards smaller metadata collections of less than 600 records.

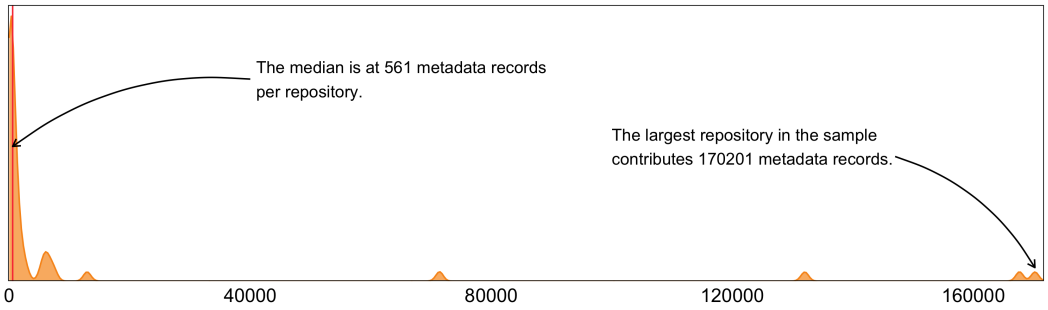

Figure 9: Number of metadata records per repository

Figure 10 shows the accumulation in the number of metadata records over the last 50 years. Some information objects described in the sample are several hundred years old, the oldest being published in 1476. However, most information objects were published in the last ten years. Some metadata records state implausible or incorrect values, notably "0000" and "9999". These values were not considered in temporal analysis. Although this issue raises questions regarding quality dimensions like correctness, this thesis mainly focuses on metadata completeness. The content of metadata elements was not taken into account.

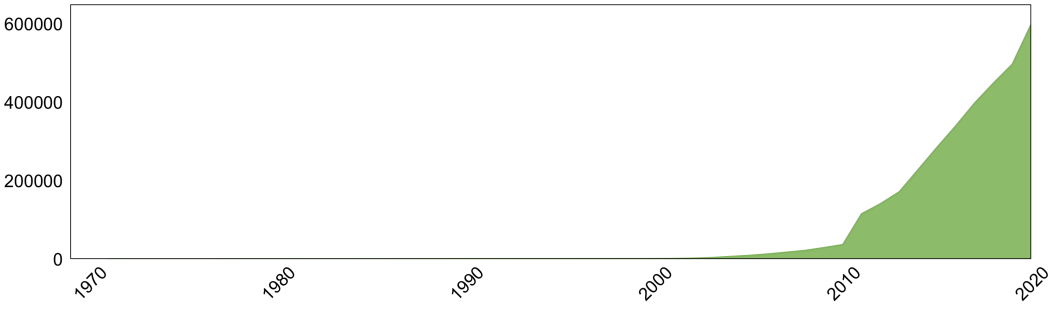

Figure 10: Cumulative number of metadata records per publication year, 1970-2020

The majority (99.82 %, n = 605006) of metadata records in the sample were registered with DataCite in the last two years (2019 and 2020). Therefore, the most common schema versions determined by approximation were 4.3 and 4.2.

3.6 Metrics for the evaluation of metadata records

This thesis focuses on the quantitative evaluation of metadata records. Therefore, not all dimensions of metadata quality described in the literature section above could be considered (Bruce and Hillmann, 2004). Included were aspects of the dimensions *completeness*, *provenance*, *logical consistency and coherence*, and *timeliness*:

Completeness
- use of schema elements
- completeness of metadata records
- comprehensiveness of descriptions
- use of persistent identifiers

Provenance
- changes to metadata records

Logical consistency and coherence
- collection homogeneity

Timeliness
- metadata timeliness

4 Findings

4.1 Properties of metadata collections

The first section of the analysis is based on RQ1 and explores characteristics of the metadata sample. A list of schema elements, their definition and obligation level can be found in the Appendix C.

4.1.1 Use of schema elements

Metadata records in the sample use between 8 and 52 metadata elements. A metadata record comprises 18.73 elements on average, with a median of 19 elements (see table 4).

	min	max	mean	median
number of elements	8	52	18.73	19

Table 4: Summary of the number of elements present per metadata record

Figure 11 shows the use of schema elements by obligation level (mandatory, recommended or optional) across all metadata records. Obligation levels of schema elements are based on the documentation of the most current version of the DataCite Metadata Schema (version 4.3). In this and all following steps of the analysis, the approximated schema version (described in the methodology section above) was considered when determining the use of schema elements, as some elements were introduced in later versions of the DataCite Metadata Schema.

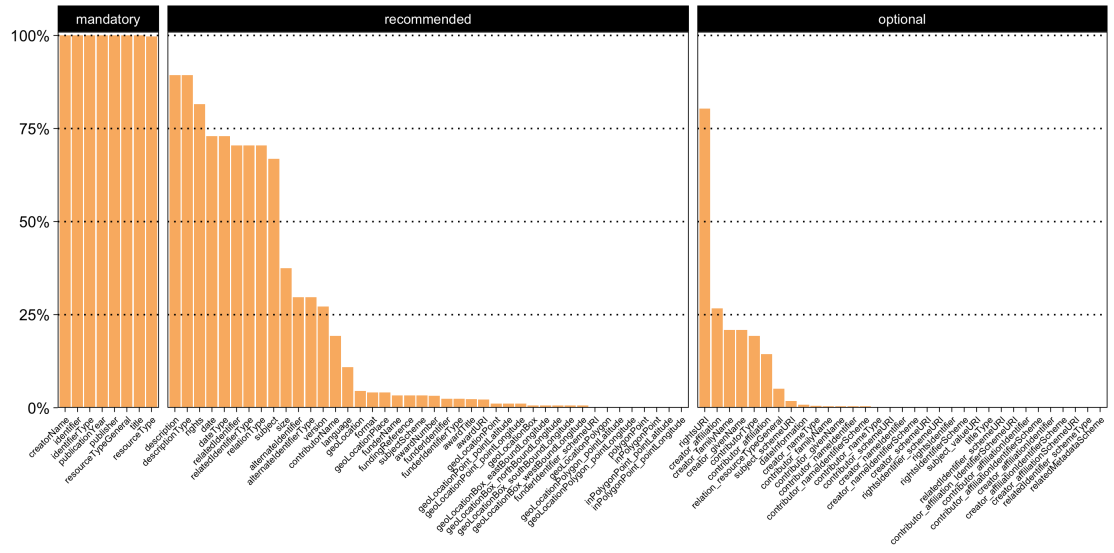

Figure 11: Use of schema elements by obligation level

The results show that of the 8 mandatory schema elements, 7 are present for all metadata records in the sample. One mandatory element (*resourceType*) is present in 99.7 % of all metadata records. Recommended elements are present less frequently overall. Of the 42 elements, 9 are used in more than 50 % of the metadata records. The elements *description* and *descriptionType* are most common in this group (89.3 %). 81.5 % of the metadata records include the element *rights*, 73.0 % *date* and *dateType*, 70.4 % *relatedIdentifier*, *relatedIdentifierType* and *relationType*, and 66.9 % *subject*. 3 recommended elements are not present in any metadata record in the sample, and an additional 15 elements are present in less than 5 % of the metadata records. The 33 optional elements in the DataCite metadata schema are overall used least frequently. The most common optional element is *rightsURI*, which is present in 80.3 % of metadata records. The other elements are used significantly less, with 19 elements being present in less than 1 % of the sample, and an additional 5 elements not being present in any metadata record.
Figure 12 shows the use of elements by metadata type. Metadata types were assigned based on the definition of NISO (Riley and Niso, 2017, p.6).

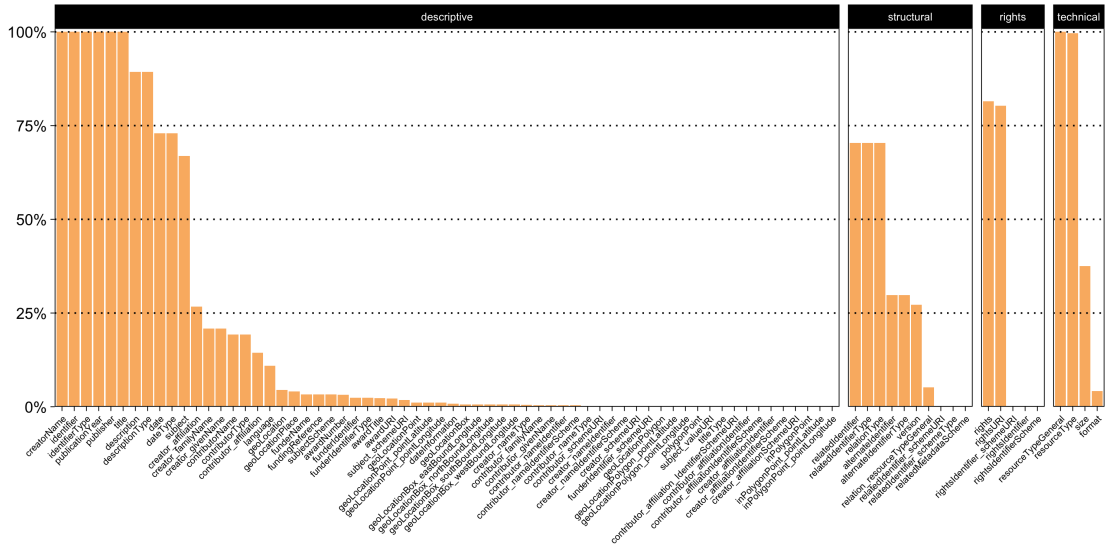

Figure 12: Use of schema elements by metadata type

Most (64) elements in the DataCite Metadata Schema are descriptive. 10 elements refer to structural, 5 to legal, and 4 to technical aspects. No element is focused specifically on long-term preservation. Mandatory elements are present in the group of structural and legal elements. On average, technical metadata is most complete (60.3 %), with rights (32.4 %), structural (30.3 %) and descriptive (18.2 %) metadata being used less frequently. Of the elements that are not used in the sample, most are categorized as descriptive metadata.

4.1.2 Completeness of metadata records

The average completeness of metadata records in the sample is 24.72 %, with the median at 25.3 %. As Table 5 shows, metadata records use between 10.84 % and 109.68 % of the elements available. For 3 metadata records, the completeness is larger than 100 %, meaning that more elements were used than available in the approximated schema version. These metadata records were likely updated after a more recent version of the DataCite Metadata Schema became available.

	min	max	mean	median
record completeness overall	10.84 %	109.68 %	24.72 %	25.3 %
average record completeness by repository	13.47 %	48.47 %	25.86 %	24.6 %

Table 5: Summary of overall record completeness and average record completeness by repository

On average, metadata records of the repositories in the sample vary between 13.47 % and 48.47 %, with an average of 25.86 % and a median of 24.6 %.

4.1.3 Collection homogeneity

To determine the homogeneity of metadata records, the most common combination of metadata elements used at least once, and the number of metadata records using this combination of elements were identified for each repository. As table 6 shows, between 9.87 % and 100 % of metadata records of a repository's collection use the same common set of metadata elements, with an average of 50.85 % and a median of 45.36 %. The set of common elements comprises between 9 and 39 elements, with an average of 19.55 and a median of 20.

	min	max	mean	median
collection homogeneity	9.87 %	100 %	50.85 %	45.36 %
size of the common element set	9	39	19.55	20

Table 6: Summary of collection homogeneity and the common element set

There is a negative relationship between the size of the common element set and collection homogeneity, meaning that collection homogeneity tends to decreases as the size of the common element set grows. The correlation between the two variables, determined by Spearman's rank correlation coefficient, is moderate (Spearman's ρ = -0.446; p = 0.002) and significant at a 5 % significance level. There is also a moderate negative correlation between the average record completeness at a repository and the homogeneity of its metadata collection (Spearman's ρ = -0.379; p = 0.009). The relationship is significant at a 5 % significance level. Figure 13 shows the relationship between the two variables.

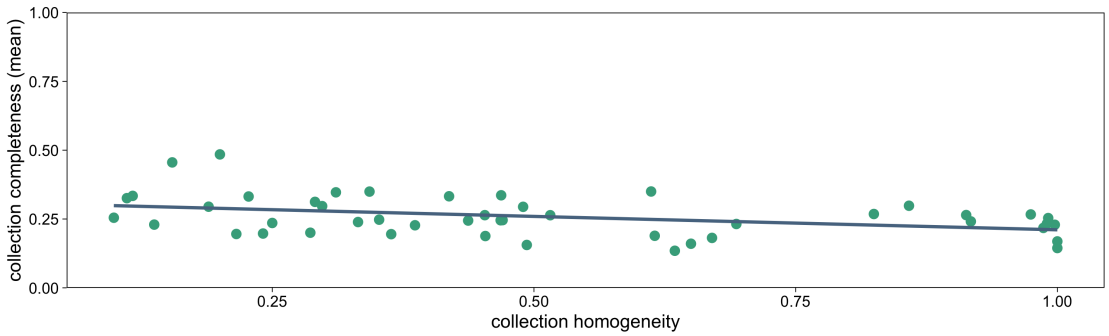

Figure 13: Correlation of collection homogeneity and average record completeness by repository

4.1.4 Comprehensiveness of descriptions

The comprehensiveness of metadata descriptions was investigated by analyzing the elements *title* and *description*. As table 7 shows, metadata records in the sample have between 1 and 3 titles. The sum of characters in all titles for each metadata record varies between 1 and 2117 characters, with an average of 79.11 and a median of 64 characters.

	min	max	mean	median
number of titles	1	3	1	1
characters in titles	1	2117	79.11	64

Table 7: Summary of number of titles and characters in titles per metadata record

Metadata records have a minimum of 0 and a maximum of 6 descriptions, as shown in table 8. The character length for all descriptions of a metadata record varies between 0 and 54468 characters, with an average of 487.3 and a median of 151 characters.

	min	max	mean	median
number of descriptions	0	6	0.98	1
characters in descriptions	0	54468	487.3	151

Table 8: Summary of number of descriptions and characters in descriptions per metadata record

4.1.5 Use of persistent identifiers

Of all persistent identifiers, related identifiers are used most frequently in the sample – 70.4 % (426749) of metadata records include a reference to a related resource. 29.7 % (180275) of metadata records specify at least one alternative identifier.
Funder identifiers are present less commonly, with 2.46 % (14570) records using this element. Name identifiers for contributors are used in 0.38 % (2302), and name identifiers for creators in 0.11 % (651) of records. Affiliation identifiers are only used in one metadata record to specify the affiliation of a contributor.
Overall, 35 repositories use related identifiers, 21 use alternative identifiers, 16 use funder identifiers, 14 use name identifiers for contributors, 4 use name identifiers for creators, and one uses affiliation identifiers.

4.1.6 Metadata timeliness

Table 9 summarizes the delay between publication year and the year of metadata creation for all metadata records in the sample, and the average by repository. Overall, between -1 and 544 years pass between a dataset is published and the metadata is registered with DataCite, with an average of 5 and a median of 4 years. The negative delay can be explained by datasets published after an embargo period, meaning the metadata was registered before the data was made available. On average, the

delay per repository varies between 0 and 67.74 years, with an average of 5.52 and a median of 1.87 years.

	min	max	mean	median
overall delay (years)	-1	544	5	4
mean delay by repository (years)	0	67.74	5.52	1.87

Table 9: Summary of delay between publication year and year of metadata creation overall and by repository (mean)

4.2 Relationship between properties of repositories and their metadata collections

In the next step of the analysis, based on RQ2, the influence of selected repository characteristics on metadata was investigated at the level of metadata elements, metadata records and metadata collections.
The repository characteristics considered for this analysis were repository type, repository subject, and certification status. The statistical tests used for describing differences between repositories with the selected characteristics require independent groups. Repositories with overlapping characteristics were therefore excluded from the analysis. In preparation for the analysis, the overlap in the selected variables was checked. There was no overlap for a repository's certification status. 6 repositories with overlapping types were omitted for this analysis (see Figure 8 above). The repositories in the sample show a large overlap in terms of assigned subject categories – 27 repositories are assigned all 4 subject categories (see Figure 7 in the method section above). Therefore, the influence of a repository's affiliation to a subject on its metadata collection could not be investigated in this thesis. After conducting an Anderson-Darling normality test, the assumption of a normal distribution for all dependent variables was rejected. Therefore, non-parametric methods were chosen over parametric methods for investigating differences across groups. As there are three levels (disciplinary, institutional and other) of the independent variable repository type, the Kruskal-Wallis test was selected (effect sizes are reported in η^2). In the case of the independent variable certification status, there are two levels (true and false), therefore, the Mann-Whitney U-test was used (effect sizes are reported in r).

4.2.1 Metadata elements

Differences in the completeness of individual metadata elements were analyzed across repository types. Table 10 shows results of the Kruskal-Wallis test for individual metadata elements that are significant at a 5 % significance level. Differences in element completeness across repository types are significant for *geoLocationPolygon*, *language* and *polygonPoint*, with a moderate effect size in all three cases.

element	η^2	p-Value
geoLocationPolygon	0.006	0.126
language	0.042	0.173
polygonPoint	0.012	0.126

Table 10: Results of the Kruskal-Wallis test (repository type) for the completeness of individual metadata elements

element	r	p-Value
contributorType	0.348	0.018
contributorName	0.348	0.018
affiliationIdentifier (contributor)	0.349	0.02
affiliationIdentifierScheme (contributor)	0.349	0.02
schemeURI (contributor affiliation identifier)	0.349	0.02
dateInformation	0.349	0.02
relatedIdentifier	0.29	0.048
relatedIdentifierType	0.29	0.048
relationType	0.29	0.048
format	0.316	0.032
geoLocation	0.51	< 0.001
geoLocationBox	0.493	< 0.001
geoLocationPlace	0.428	0.004

Table 11: Results of the Mann-Whitney U-test (certification status) for the completeness of individual metadata elements

At the level of individual metadata elements, differences in completeness were also analyzed across repositories with and without formal certification. The results of the Mann-Whitney U-test for individual metadata elements that are significant at a 5 % significance level are displayed in table 11. The effect sizes are small for the elements *relatedIdentifier*, *relatedIdentifierType* and *relationType*. Effect sizes for the elements *contributorType*, *contributorName*, *affiliationIdentifier* (contributor), *affiliationIdentifierScheme* (contributor), *schemeURI* (contributor affiliation identifier), *dateInformation*, *format*, *geoLocationBox*, and *geoLocationPlace* are moderate. The effect size for the element *geoLocation* is large.

4.2.2 Metadata records

The boxplot in figure 14 shows the variance in dependent variables describing metadata records (record completeness, description length and time lag) across repository types.

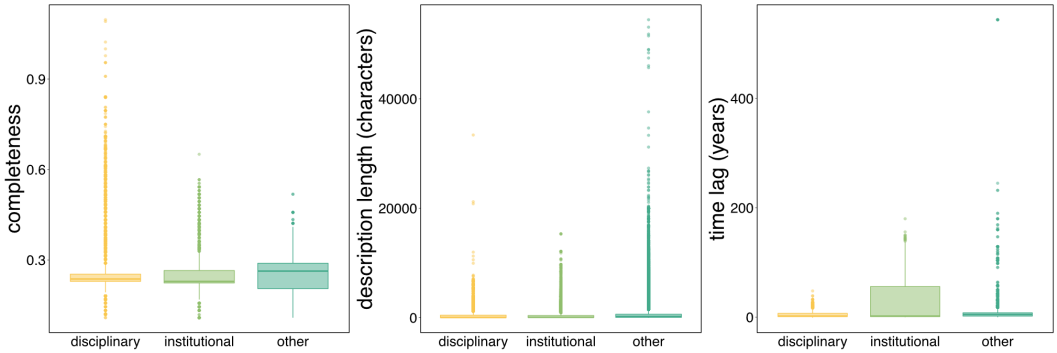

Figure 14: Boxplots of variables describing metadata records by repository type

The average record completeness is highest for disciplinary repositories (26.1 %), followed by repositories of the type *other* (24.8 %) and institutional repositories (24.5 %). Descriptions are on average most detailed for repositories of the type *other* (556.68 characters), and shorter for institutional (468.5 characters) and disciplinary (466.94 characters) repositories. Disciplinary repositories show the shortest delay in the availability of metadata records after data publication (4.14 years), with repositories of the type *other* taking slightly longer (5.06 years). On average, institutional repositories publish metadata records 22.07 years after datasets are made available.

variable	η^2	p-Value
record completeness	0.006	< 0.001
description length	0.012	< 0.001
time lag	0.011	< 0.001

Table 12: Results of the Kruskal-Wallis test (repository type) for variables describing metadata records

The results of the Kruskal-Wallis test summarized in table 12 show that differences in all selected variables across repository types are significant at a 5 % significance level. Effect sizes of the differences in record completeness, description length and time lag are small, however.
Variance in dependent variables describing metadata records (record completeness, description length and time lag) across certification status is displayed in figure 15.

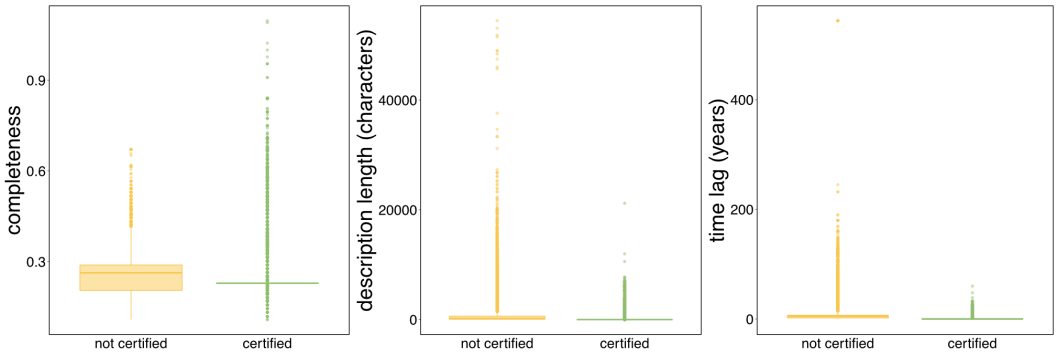

Figure 15: Boxplots of variables describing metadata records by certification status

On average, repositories without formal certification offer metadata records with a slightly higher degree of completeness (24.8 %) compared to repositories with formal certification (24.5 %). At 549.31 characters on average, descriptions are longer at repositories without formal certification than at repositories with formal certification (185.69 characters). Repositories with formal certification make metadata available on average 1.36 years after datasets are published, whereas repositories without formal certification publish metadata on average 5.75 years after datasets.

variable	r	p-Value
record completeness	0.145	< 0.001
description length	0.322	< 0.001
time lag	0.459	< 0.001

Table 13: Results of the Mann-Whitney U-test (certification status) for variables describing metadata records

Table 13 shows the results of the Mann-Whitney U-test for variables describing metadata records. All differences are significant at a 5 % significance level. In the case of certification status, effect sizes of differences across groups are moderate for description length and time lag, and small for record completeness.

4.2.3 Metadata collections

The variable collection homogeneity describes metadata at the level of repository collections. Figure 16 shows boxplots of collection homogeneity across repository type (A) and certification status (B).

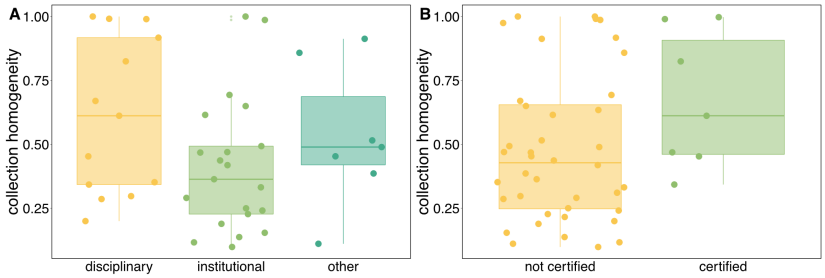

Figure 16: Boxplots of variables describing metadata collections by repository type (A) and certification status (B)

On average, metadata collections of disciplinary repositories are most homogenous (61.1 %), followed by repositories of the type *other* (53.2 %) and institutional repositories (41.1 %). Repositories with formal certification have on average more homogenous metadata collections (67 %) compared to repositories without formal certification (48 %). Differences in collection homogeneity are neither significant across repository types nor across repositories with and without formal certification (at a 5 % significance level).

4.3 Changes to metadata records

In the last step of the analysis, following RQ3, changes to metadata records are analyzed.
Information on the history of metadata records is made available by DataCite for all metadata records registered since March 10th, 2019. Included in the analysis were metadata records registered after that date for which a metadata version was available. As a result, 164775 datasets from 32 repositories were analyzed. The statistical tests used to analyze differences across repository types and certification status are identical to the tests used in the section above.
As table 14 shows, metadata records in the sample were changed between 0 and 1053 times, with an average of 3.72 and a median of 2 changes. On average by repository, metadata records are changed between 0 and 85.27 times, with an average of 4.35 and a median of 0.93 changes.

	min	max	mean	median
number of changes to metadata records	0	1053	3.72	2
mean number of changes to metadata records by repository	0	85.27	4.35	0.93

Table 14: Summary of the number of changes to metadata records

Overall, 73.9 % (446991) of metadata records were changed. The percentage of changed metadata records by repository varies between 0.8 % and 100 %, with an average of 54.21 % and a median of 48.38 %.

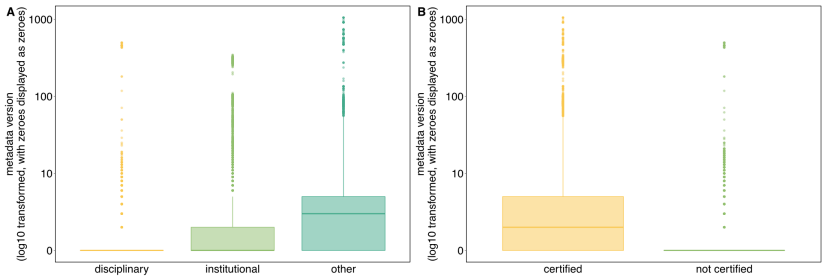

Figure 17: Boxplots of changes to metadata records by repository type (A) and certification status (B)

An analysis of differences in the number of changes to metadata records across repository types shows that on average, metadata records are changed most frequently at institutional repositories (8.36 changes), and less frequently at repositories of the type *other* (4.22 changes) and disciplinary repositories (1.07 changes). Differences in the number of changes to metadata records across repository types are displayed in figure 17 (A). Results of a Kruskal-Wallis test show that these differences across repository types are significant at a 5 % significance level, with a moderate effect size ($\eta^2 = 0.111$; $p < 0.001$).
On average, metadata records are changed more often at repositories with formal certification (4.29 changes) compared to repositories without a certificate (0.95 changes). Differences in the number of changes to metadata records across certification status are displayed in figure 17 (B).
These differences across repositories with and without formal certification are significant at a 5 % significance level with a moderate effect size ($r = 0.466$; $p < 0.001$), as the results of a Mann-Whitney U-test demonstrate.

5 Discussion

Methodology

This thesis describes a reliable method for matching repository descriptions in re3data and DataCite metadata records, enabling a joint analysis of the two data sources. A persistent identifier for research data repositories used by both sources would facilitate metadata aggregation and bibliometric analysis, aside from other advantages.
The matching process shed light on serious concerns regarding the global repository landscape. Particularly noticeable is the geographical distribution of institutions associated with the repositories the sample, which is strongly skewed towards Europe and North America. A likely explanation for this observation is that the sampling strategy set high standards for the technical maturity of the selected repositories. The uneven geographical representation highlights a globally imbalanced distribution of resources, where not all repositories can expose their metadata via standardized interfaces or are able to assign DOIs for datasets. Collection sizes of repositories in the sample are also indicative of centralization tendencies in data sharing. Most collection sizes comprised less than 600 records, whereas figshare, Zenodo and DRYAD surpassed 100000 records. These centralization tendencies should be examined more closely in future research, since they have serious implications on the idea of a global *data commons*, where anyone can participate in and contribute to data sharing and reuse.
In general, the large number of repositories without interfaces was concerning. The matching process demonstrated that only few repositories offer OAI-PMH interfaces to their metadata collections, and even if the information in re3data indicates a repository has an OAI-PMH interface, it might not be operational anymore. This issue is not limited to OAI-PMH interfaces: as of November 20th 2020, 54.61 % (1416) of the 2593 repositories listed in re3data did not offer an interface of any type. This lack of APIs significantly reduces the visibility of research data and has serious implications for the implementation of the FAIR Principles. For example, the FAIR indicator "F4: Metadata is offered in such a way that can be harvested and indexed" is rated *essential* the RDA FAIR Data Maturity Model Working Group (Research Data Alliance FAIR Data Maturity Model Working Group, 2020; p. 11).
The analysis centered around RQ2 examined the relationship between characteristics of repositories and their metadata collections. It is important to note that the statistical tests used do not imply causality, as there likely are confounding factors besides the independent variables *repository type* and *certification status* that influence the dependent variables. One likely confounding factor that could not be considered are the resources available to a repository, for example repository staff. Future research could clarify the influence of this factor on metadata collections.

Use of schema elements

The analysis first explored properties characterizing metadata for research data in general (RQ1). On average, the metadata records analyzed in this thesis used 18.73 metadata elements. This exceeds

the 8 mandatory elements in the current version of the DataCite Metadata Schema (4.3), meaning that metadata descriptions on average are more detailed than the required minimum. No metadata record uses more than 52 elements, although the current version of the DataCite Metadata Schema (4.3) defines 83 elements in total.

Not surprisingly, the analysis of the use of schema elements showed that of the elements defined by the DataCite Metadata Schema, mandatory elements are used most frequently. Mandatory elements are required upon metadata registration, and with the exception of *resourceType*, which is used in 99.7 % of records, all mandatory elements are present in all metadata records. Why *resourceType* is not available for all metadata records remains unclear. Compared to mandatory elements, recommended elements are used less, but are present more frequently than optional elements. This indicates that the obligation level of the metadata schema contributes to metadata completeness, and that recommending the use of certain schema elements has an effect on its use. Besides the mandatory elements, most metadata records include descriptions of the dataset (*description* and *descriptionType*) and information on legal aspects (*rights* and *rightsURI*). These elements are used in more than 75 % of the metadata records. Other commonly used metadata elements that are used in more than 50 % of the metadata records provide information on dates (*date* and *dateType*), related resources (*relatedIdentifier*, *relatedIdentifierType*, and *relationType*), and subjects (*subject*). Surprisingly, some recommended elements providing essential practical information for potential data users are less common. Information on the *size* of a dataset is present in 37.5 % of metadata records, and a data *version* is stated in 27.2 % cases. The *format* is explicitly mentioned in only 4.16 % of the metadata records. Missing information in these elements can have negative effects on the reuse of the datasets described.

Touching on RQ2, differences in the use of schema elements were analyzed across repository types and certification status. In the case of the element *format*, the frequency of use is significantly higher at repositories with formal certification compared to repositories without a certificate. The documentation of the certification scheme CoreTrustSeal mentions the format of datasets explicitly in the requirements Appraisal (R8) and Data Reuse (R14) (CoreTrustSeal Standards and Certification Board, 2019). Repositories certified by the CoreTrustSeal may therefore be more aware of the benefits of providing information on formats. Automated metadata creation could be considered to retrospectively complete metadata records by adding technical specifications inherent to datasets, like size and format.

The element *contributorName* is used in only 19.3 % of the metadata records. In the documentation of the DataCite Metadata Schema, contributors are defined as "The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource." (DataCite Metadata Working Group, 2019; p. 18) The low prevalence of information on contributors is interesting, particularly in the context of discussions regarding the invisibility of metadata labor and crediting data stewards' contributions to managing research data. As the analysis based on RQ2 demonstrated, certified repositories are significantly more likely to specify contributors than repositories without formal certification. A possible explanation for these differences could be that certified repositories are more aware of the value of data stewards' labor, or the responsible institution seeks recognition for its contributions.

Analyzing element use across NISO metadata types highlighted the lack of elements in the DataCite Metadata Schema regarding the long-term preservation of research data. This is not surprising, however, as DataCite primarily focuses on assigning persistent identifiers and making datasets discoverable. On average, technical metadata is most complete, although essential information are underused, in particular size and format, as discussed above. Rights metadata is also not available for all metadata records. This is concerning, since it means that the legal parameters of using these datasets remain unclear to potential reusers, posing a significant barrier to reuse. To resolve this issue, making rights information mandatory should be considered for future versions of the DataCite Metadata Schema.

Use of persistent identifiers

Another area where information is missing from metadata records are persistent identifiers. As the descriptive analysis based on RQ1 demonstrated, persistent identifiers are underused, in particular name identifiers. Name identifiers apply to all records in the metadata sample, since specifying the creator(s), the main person(s) involved in creating a dataset, is mandatory in the DataCite Metadata Schema. An interesting observation is that name identifiers are used more frequently for contributors than for creators. A possible explanation for this variance in the use of name identifiers could be that data stewards are more aware of the benefits of name identifiers and have adopted their use at a higher rate compared to data creators. Affiliation identifiers are also underused, although they were only introduced to the DataCite Metadata Schema in the most recent version. In the sample, affiliation identifiers are used in only one metadata record. Therefore, the analysis based on RQ2 of the difference in the use of affiliation identifiers for contributors across certification status is not informative, even though the difference is significant. Persistent identifiers referring to related sources are used more frequently, in 70.4 % (426749) of the metadata records. These identifiers likely resolve to journal publications based on the datasets. Journals increasingly publish data guidelines that ask for references to the reported data. Authors also potentially benefit from establishing links between text and data publications, in the form of increased citation rates (Colavizza et al., 2020). In the sample, related identifiers were significantly more likely to be present in metadata records of repositories with formal certification, as an analysis based on RQ2 showed.
It is important to note that related identifiers and alternative identifiers, which are present in 29.7 % (425816) of the metadata records in the sample, can also be used for versioning dynamic or changing datasets. The purpose of these references can not be investigated in the context of this thesis.

Completeness of metadata records

Following RQ1, metadata completeness first appears to be quite low at the level of metadata records, with only 24.72 % of the available elements being used on average. Collection completeness, the average completeness of metadata records at a given repository, varies notably between 13.47 % and 48.47 %. A likely explanation for the low metadata completeness is that not all elements in the DataCite Metadata Schema are applicable to all datasets. As discussed in the literature section above, not all statements are useful for describing a given dataset. A good example is geolocation information: the DataCite Metadata Schema offers several elements for describing geolocation, but not all datasets are associated with a region or place. If descriptions of geolocation are not applicable to a dataset, the completeness of the metadata record describing this dataset is lower compared to other records. The observed variations in completeness across repositories could be explained by the characteristics of collections the repositories hold. For example, a repository collecting geoscience datasets with associations to specific locations may achieve higher metadata completeness compared to other repositories. Metadata completeness therefore is not just an indicator of how well a dataset is described, but also of how well a metadata schema is suited for describing the datasets. The seemingly low metadata completeness is in part a result of using a generic metadata schema for describing diverse datasets.
Differences in record completeness are significant across repository types and certification status, but the effect sizes are small, as an analysis followind RQ2 demonstrated. Subsequent studies could determine whether subject areas or disciplines lead to larger differences in the completeness of metadata records.

Collection homogeneity

In the analysis based on RQ1, metadata collections were inspected with regard to collection homogeneity. On average, 50.85 % of metadata records at a given repository use the most common combination of metadata elements. This can be interpreted as a potential indicator of consistent practices at repositories for describing data. For two repositories, the collection homogeneity is 100 %, however, the common element set is small in these cases (11 and 13 elements). The common element set comprises 19.55 elements on average but varies considerably between 9 and 39 elements. There is a significant negative relationship between the size of the common element set and collection homogeneity, meaning that the homogeneity of metadata decreases as the common element set grows in size. A related observation is the correlation of collection completeness and collection homogeneity. Both observations indicate that collection homogeneity decreases as descriptions get more comprehensive. This can also be explained, at least in part, by the generic nature of the DataCite Metadata Schema. Not all elements in the element set are suitable for the description of all datasets in a collection, therefore more detailed descriptions may reduce the homogeneity of the collection overall. Homogeneity is therefore likely also affected by subjects, which could not be examined in this thesis.
In any case, to describe results concerning RQ2, differences in collection homogeneity between repository types and certification status are not significant.

Comprehensiveness of descriptions

Contributing to RQ1, length of titles and descriptions were analyzed as a proxy for the comprehensiveness of metadata descriptions. On average, titles are 79.11 characters long. The element *title* is mandatory, but the titles of 44 metadata records comprise only one character, effectively circumventing the obligation level. Going forward, DataCite could specify additional restrictions and validate metadata upon DOI registration to prevent metadata records without meaningful titles. The element *description* is recommended by DataCite, and 10.7 % (64817) of metadata records in the sample do not include a description. On average, descriptions comprise 487.3 characters.
An analysis based on RQ2 demonstrated that differences in description length are significant across repository types and certification status. Effect sizes are small for repository types and moderate for certification status. Descriptions are on average most detailed for repositories of the type *other* and for repositories without formal certification. Both observations are notable, because it could be assumed that certified repositories or repositories with a disciplinary focus have more resources available and are therefore more likely to provide comprehensive descriptions of datasets. Descriptions of datasets could be examined more closely in subsequent studies, for example to determine who produces these summaries and how. Furthermore, the content of descriptions could be analyzed using text mining.

Metadata timeliness

The timeliness of metadata, one aspect of RQ1, was evaluated by analyzing the time passed between the year a dataset was published and the year the metadata record describing the dataset was made available. On average, 5.52 years passed before a metadata record was registered with DataCite. A likely explanation for this metadata delay is that repositories retrospectively assign DOIs for all datasets in their collection, including those that have been published several years ago. The delays in metadata registration therefore probably reflect the age of the collection and the relative point in

time at which the repository started assigning DOIs.
The metadata delay is smallest for disciplinary repositories and for repositories with formal certification. These differences are significant, with a small effect size for repository type and a moderate effect size for certification status, as the analysis based on RQ2 showed.

Changes to metadata records

The analysis based on RQ3 showed that at some repositories, metadata records are treated as dynamic objects. Of all metadata records in the sample, 73.9 % metadata records have been changed at some point. On average, a metadata record was changed 3.72 times. However, there are notable differences in the rate of changes across individual repositories. At 11 repositories, more than 90 % of metadata records have been changed, whereas less than 1 % of metadata records have been changed at 2 repositories. Differences in metadata changes are significant across repository types, with a moderate effect size. Metadata at institutional repositories changed most frequently, and least frequently at disciplinary repositories. A possible explanation for this could be that institutional repositories tend to divide metadata labor among several persons, for example data providers and data stewards, and publish changes as the results of an iterative process. This question could be addressed in subsequent studies. Similar assumptions could be made for certified repositories, since significant differences in the number of changes to metadata records were also observed across certification status, with certified repositories being more likely to change metadata records. The effect size was moderate.

6 Conclusion

In conducting a quantitative assessment informed by metadata quality requirements, this thesis provides an overview of metadata for research data, including the interrelation between repository characteristics and their metadata (collections).

Following RQ1, characteristics of metadata for research data were analyzed. Overall, obligation levels of schema elements have an impact on their use. Some schema elements are underused considering their contribution to facilitating data (re-)use, for example rights information. In future revisions of the DataCite Metadata Schema, obligation levels of these elements could be adapted to encourage their use, thereby facilitating data (re-)use. Completeness of metadata records vary across repositories, which could be an indicator for distinct metadata practices at individual research data repositories, but is likely also skewed by using a generic metadata schema for describing diverse datasets. Within repositories, metadata descriptions are relatively homogenous, suggesting that repositories have developed consistent practices for describing data. As metadata descriptions get more detailed, metadata homogeneity decreases. On average, descriptions comprise 487.3 characters, and 5.52 years passed between the year a dataset was published and the metadata record was registered with DataCite.

RQ2 focused on the relationship between characteristics of repositories (repository type and certification status) and metadata. Differences in the completeness of metadata records, description length and timeliness were significant across repository types and certification status. Effect sizes of differences in description length and timeliness were larger across certification status than repository types, suggesting that formal certification has a stronger influence on these metrics. Differences in collection homogeneity were neither significant across repository type nor certification status.

Following RQ3, changes to metadata records were analyzed. Overall, most metadata records in the sample were changed, which supports the conceptualization of metadata for research data as dynamic and changeable objects. There are notable differences in the rate of changes across individual repositories, and these differences are significant across repository types. Metadata were changed most frequently at institutional repositories, which could be an indicator of distinct practices for describing datasets, for example consistent workflows with division of labor.

In giving an overview of metadata for research data, this thesis contributes to understand the effects of data stewardship.

7 Limitations

The process for matching the two data sources used in this thesis requires high technological maturity of repositories. Therefore, the number of repositories included in the analysis is limited, and results should not be considered representative of all research data repositories.
Standardization of records in the DataCite Metadata Store is still limited, potentially restricting its applicability in scientometric analyses (Robinson-Garcia et al., 2017).
To refine the analysis of the use of metadata elements, the schema version was approximated based on the date a metadata record was first registered with DataCite. Considering the high rate of changed metadata records, this approach is not ideal, and an an approximation based on the last update of a metadata record would be more precise. Therefore, metadata completeness reported in this thesis likely tends to be too high.
Metadata comprehensiveness was approximated by the length of descriptions. Factors relating to the content of descriptions were not considered in this paper.

8 References

Ackoff, Russell L. (1989). „From data to wisdom“. In: *Journal of Applied Systems Analysis* 16, pp. 3–9.

Assante, Massimiliano et al. (2016). „Are Scientific Data Repositories Coping with Research Data Publishing?“ In: *Data Science Journal* 15. DOI: 10.5334/dsj-2016-006.

Baca, Murtha (2016). „Practical Principles for Metadata Creation and Maintenance“. In: *Introduction to Metadata*. Ed. by Murtha Baca. URL: http://www.getty.edu/publications/intrometadata (visited on 12/06/2020).

Baker, Karen S. and Matthew S. Mayernik (2020). „Disentangling knowledge production and data production“. In: *Ecosphere* 11.7, e03191. DOI: 10.1002/ecs2.3191.

Ball, Alexander et al. (2014). „Building a Disciplinary Metadata Standards Directory“. In: *International Journal of Digital Curation* 9.1, pp. 142–151. DOI: 10.2218/ijdc.v9i1.308.

Bates, Jo (2018). „The politics of data friction“. In: *Journal of Documentation* 74.2, pp. 412–429. DOI: 10.1108/JD-05-2017-0080.

Benjelloun, Omar, Shiyu Chen, and Natasha Noy (2020). „Google Dataset Search by the Numbers“. In: *The Semantic Web – ISWC 2020*. Ed. by Jeff Z. Pan et al. Berlin: Springer, pp. 667–682. DOI: 10.1007/978-3-030-62466-8_41.

Bettivia, Rhiannon S. (2016). „The power of imaginary users: Designated communities in the OAIS reference model“. In: *Proceedings of the Association for Information Science and Technology* 53.1. Wiley, pp. 1–9. DOI: 10.1002/pra2.2016.14505301038.

Borgman, Christine L (2016). *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA; London: The MIT Press.

Bowker, Geoffrey C. and Susan Leigh Star (2000). *Sorting things out: classification and its consequences*. Inside technology. Cambridge, MA; London: The MIT Press.

Bowker, Geoffrey C. et al. (2010). „Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment“. In: *International Handbook of Internet Research*. Ed. by Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen. Berlin: Springer, pp. 97–117. DOI: 10.1007/978-1-4020-9789-8_5.

Bruce, Thomas R. and Diane I. Hillmann (2004). *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*. URL: https://hdl.handle.net/1813/7895 (visited on 10/01/2020).

Bugaje, Maryam and Gobinda Chowdhury (2017). „Is Data Retrieval Different from Text Retrieval? An Exploratory Study“. In: *Digital Libraries: Data, Information, and Knowledge for Digital Lives*. Ed. by Songphan Choemprayong, Fabio Crestani, and Sally Jo Cunningham. Berlin: Springer, pp. 97–103. DOI: 10.1007/978-3-319-70232-2_8.

Candela, Leonardo et al. (2015). „Data journals: A survey“. In: *Journal of the Association for Information Science and Technology* 66.9, pp. 1747–1762. DOI: 10.1002/asi.23358.

CASRAI. *CASRAI Research Data Management Glossary: Repository*. URL: https://casrai-test.evision.ca/glossary-term/repository/ (visited on 12/06/2020).

CCSDS Secretariat (2012). *Reference Model For An Open Archival Information System (OAIS)*. URL: https://public.ccsds.org/Pubs/650x0m2.pdf (visited on 12/06/2020).

Chang, Hasok (2005). „A Case for Old-Fashioned Observability, and a Reconstructed Constructive Empiricism". In: *Philosophy of Science* 72.5, pp. 876–887. DOI: 10.1086/508116.

Chapman, Adriane et al. (2019). „Dataset search: a survey". In: *The VLDB Journal* 29, 251–272. DOI: 10.1007/s00778-019-00564-x.

CODATA-ICSTI Task Group on Data Citation Standards and PractOut of Cite (2013). „Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data". In: *Data Science Journal* 12. DOI: 10.2481/dsj.OSOM13-043.

Colavizza, Giovanni et al. (2020). „The citation advantage of linking publications to research data". In: *PLOS ONE* 15.4, e0230416. DOI: 10.1371/journal.pone.0230416.

Contreras, Jorge L. (2010). *Bermuda's Legacy: Policy, Patents and the Design of the Genome Commons*. SSRN Scholarly Paper ID 1667659. Rochester, NY: Social Science Research Network. URL: https://papers.ssrn.com/abstract=1667659 (visited on 12/06/2020).

CoreTrustSeal Standards and Certification Board (2019). *CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022*. DOI: 10.5281/zenodo.3638211.

Cousijn, Helena et al. (2019). „Bringing Citations and Usage Metrics Together to Make Data Count". In: *Data Science Journal* 18. DOI: 10.5334/dsj-2019-009.

DataCite. *DataCite - Members*. URL: https://datacite.org/members.html (visited on 12/06/2020).

– *DataCite's Value*. URL: https://datacite.org/value.html (visited on 12/06/2020).

DataCite Metadata Schema. DataCite. URL: https://schema.datacite.org/ (visited on 12/06/2020).

DataCite Metadata Working Group (2019). „DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.3". In: in collab. with Madeleine de Smaele et al. DOI: 10.14454/7XQ3-ZF69.

Donaldson, Devan Ray, Ewa Zegler-Poleska, and Lynn Yarmey (2020). „Data managers' perspectives on OAIS designated communities and the FAIR principles: mediation, tools and conceptual models". In: *Journal of Documentation* 76.6, pp. 1261–1277. DOI: 10.1108/JD-10-2019-0204.

Downey, Greg, Kristin R. Eschenfelder, and Kalpana Shankar (2019). „Talking About Metadata Labor: Social Science Data Archives, Professional Data Librarians, and the Founding of IASSIST". In: *Historical Studies in Computing, Information, and Society*. Ed. by William Aspray. Berlin: Springer, pp. 83–113. DOI: 10.1007/978-3-030-18955-6_5.

Edwards, Paul N. (2003). „Infrastructure and Modernity: Scales of Force, Time, and Social Organization in the History of Sociotechnical Systems". In: *Modernity and technology*. Ed. by Thomas J. Misa, Philip Brey, and Andrew Feenberg. Cambridge, MA; London: The MIT Press.

– (2013). *A vast machine: computer models, climate data, and the politics of global warming*. Cambridge, MA; London: The MIT Press.

Edwards, Paul N. et al. (2011). „Science friction: Data, metadata, and collaboration“. In: *Social Studies of Science* 41.5, pp. 667–690. DOI: 10.1177/0306312711413314.

Faniel, Ixchel M., Rebecca D. Frank, and Elizabeth Yakel (2019). „Context from the data reuser's point of view“. In: *Journal of Documentation* 75.6, pp. 1274–1297. DOI: 10.1108/JD-08-2018-0133.

Faniel, Ixchel M. and Elizabeth Yakel (2011). „Significant Properties as Contextual Metadata“. In: *Journal of Library Metadata* 11.3, pp. 155–165. DOI: 10.1080/19386389.2011.629959.

Fecher, Benedikt and Sascha Friesike (2014). „Open Science: One Term, Five Schools of Thought“. In: *Opening Science*. Ed. by Sönke Bartling and Sascha Friesike. Berlin: Springer, pp. 17–47. DOI: 10.1007/978-3-319-00026-8_2.

Gilliland, Anne J. (2008). „Setting the stage“. In: *Introduction to metadata*. Ed. by Murtha Baca and Getty Research Institute. Los Angeles, CA: Getty Research Institute.

Gitelman, Lisa and Virginia Jackson (2013). „Introduction“. In: *"Raw data" is an oxymoron*. Ed. by Lisa Gitelman. Cambridge, MA; London: The MIT Press, pp. 1–14.

Gonçalves, Rafael S. and Mark A. Musen (2019). „The variable quality of metadata about biological samples used in biomedical experiments“. In: *Scientific Data* 6.1, pp. 1–15. DOI: 10.1038/sdata.2019.21.

Google Developers: Dataset. Google Developers. URL: https://developers.google.com/search/docs/data-types/dataset (visited on 12/06/2020).

Gregg, Will et al. (2019). „A literature review of scholarly communications metadata“. In: *Research Ideas and Outcomes* 5, e38698. DOI: 10.3897/rio.5.e38698.

Gregory, Kathleen et al. (2019). „Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines“. In: *Journal of the Association for Information Science and Technology* 70.5, pp. 419–432. DOI: 10.1002/asi.24165.

Gregory, Kathleen et al. (2020a). „Lost or Found? Discovering Data Needed for Research“. In: *Harvard Data Science Review* 2.2. DOI: 10.1162/99608f92.e38165eb.

Gregory, Kathleen M et al. (2020b). „Understanding data search as a socio-technical practice“. In: *Journal of Information Science* 46.4, pp. 459–475. DOI: 10.1177/0165551519837182.

Habermann, Ted (2018a). „Metadata Life Cycles, Use Cases and Hierarchies“. In: *Geosciences* 8.5. DOI: 10.3390/geosciences8050179.

– (2018b). *Metrics for Characterizing Metadata Collections*. DataCite Blog. URL: https://blog.datacite.org/metrics-for-metadata/ (visited on 12/06/2020).

Hey, Anthony J. G., Stewart Tansley, and Kristin Tolle, eds. (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Hider, Philip (2018). *Information Resource Description: Creating and Managing Metadata*. London: Facet. DOI: 10.29085/9781783302253.

Imker, Heidi J. (2020). „Who Bears the Burden of Long-Lived Molecular Biology Databases?“ In: *Data Science Journal* 19.1, p. 8. DOI: 10.5334/dsj-2020-008.

International Organization for Standardization (2015). *Quality management systems - Fundamentals and vocabulary (ISO 9000:2015)*. URL: https://www.iso.org/standard/45481.html (visited on 12/06/2020).

Joo, Soohyung, Darra Hofman, and Youngseek Kim (2019). „Investigation of challenges in academic institutional repositories: A survey of academic librarians“. In: *Library Hi Tech* 37.3, pp. 525–548. DOI: 10.1108/LHT-12-2017-0266.

Kacprzak, Emilia et al. (2019). „Characterising dataset search—An analysis of search logs and data requests“. In: *Journal of Web Semantics* 55, pp. 37–55. DOI: 10.1016/j.websem.2018.11.003.

Kim, Jihyun, Elizabeth Yakel, and Ixchel M. Faniel (2019). „Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition“. In: *College & Research Libraries* 80.6, pp. 843–875. DOI: 10.5860/crl.80.6.843.

Kim, Youngseek and Jeffrey M. Stanton (2016). „Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis“. In: *Journal of the Association for Information Science and Technology* 67.4, pp. 776–799. DOI: https://doi.org/10.1002/asi.23424.

Kindling, Maxi et al. (2017). „The Landscape of Research Data Repositories in 2015: A re3data Analysis“. In: *D-Lib Magazine* 23.3. DOI: 10.1045/march2017-kindling.

Kitchin, Rob and Gavin McArdle (2016). „What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets:“ in: *Big Data & Society* 3.1. DOI: 10.1177/2053951716631130.

Klump, Jens (2017). „Data as Social Capital and the Gift Culture in Research“. In: *Data Science Journal* 16. DOI: 10.5334/dsj-2017-014.

Klump, Jens, Robert Huber, and Michael Diepenbroek (2015). „DOI for geoscience data - how early practices shape present perceptions“. In: *Earth Science Informatics* 9, pp. 123–136. DOI: 10.1007/s12145-015-0231-5.

Koesten, Laura et al. (2020a). *Dataset Reuse: Translating Principles to Practice*. SSRN Scholarly Paper ID 3589836. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3589836.

Koesten, Laura et al. (2020b). „Everything you always wanted to know about a dataset: Studies in data summarisation“. In: *International Journal of Human-Computer Studies* 135, p. 102367. DOI: 10.1016/j.ijhcs.2019.10.004.

Koesten, Laura M. et al. (2017). „The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour“. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. 2017 CHI Conference. Denver, CO: ACM Press, pp. 1277–1289. DOI: 10.1145/3025453.3025838.

Late, Elina and Jaana Kekäläinen (2020). „Use and users of a social science research data archive“. In: *PLOS ONE* 15.8, e0233455. DOI: 10.1371/journal.pone.0233455.

Lee, Dong Joon and Besiki Stvilia (2017). „Practices of research data curation in institutional repositories: A qualitative view from repository staff“. In: *PLOS ONE* 12.3, e0173987. DOI: 10.1371/journal.pone.0173987.

Leonelli, Sabina (2015). „What Counts as Scientific Data? A Relational Framework“. In: *Philosophy of Science* 82.5. DOI: 10.1086/684083.

– (2016). *Data-centric biology: a philosophical study*. Chicago, IL ; London: The University of Chicago Press.

– (2019). „Data — from objects to assets“. In: *Nature* 574, pp. 317–320. DOI: 10.1038/d41586-019-03062-w.
– (2020a). „Learning from Data Journeys“. In: *Data Journeys in the Sciences*. Ed. by Sabina Leonelli and Niccolò Tempini. Berlin: Springer, pp. 1–24. DOI: 10.1007/978-3-030-37177-7_1.
– (2020b). „Scientific Research and Big Data“. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2020/entries/science-big-data/ (visited on 12/06/2020).
Lin, Dawei et al. (2020). „The TRUST Principles for digital repositories“. In: *Scientific Data* 7.144. DOI: 10.1038/s41597-020-0486-7.
Loukissas, Yanni A. (2019). *All data are local: thinking critically in a data-driven society*. Cambridge, MA; London: The MIT Press.
Löffler, Felicitas et al. (2020). „Dataset Search In Biodiversity Research: Do Metadata In Data Repositories Reflect Scholarly Information Needs?“ In: *arXiv*. arXiv: 2002.12021. URL: http://arxiv.org/abs/2002.12021 (visited on 12/06/2020).
Manninen, Lauren (2018). „Describing Data: A Review of Metadata for Datasets in the Digital Commons Institutional Repository Platform: Problems and Recommendations“. In: *Journal of Library Metadata* 18.1, pp. 1–11. DOI: 10.1080/19386389.2018.1454379.
Mayernik, Matthew S. (2015). „Research data and metadata curation as institutional issues“. In: *Journal of the Association for Information Science and Technology* 67.4, pp. 973–993. DOI: 10.1002/asi.23425.
Noy, Natasha (2018). *Making it easier to discover datasets*. The Keyword. URL: https://blog.google/products/search/making-it-easier-discover-datasets/ (visited on 12/06/2020).
– (2020). *An Analysis of Online Datasets Using Dataset Search (Published, in Part, as a Dataset)*. Google AI Blog. URL: http://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html (visited on 12/06/2020).
Online Etymology Dictionary. *Online Etymology Dictionary: data*. URL: https://www.etymonline.com/word/data (visited on 12/06/2020).
Palavitsinis, Nikos (2013). „Metadata Quality Issues in Learning Repositories“. PhD thesis. Madrid: Universidad de Alcalá. URL: https://core.ac.uk/download/pdf/58910780.pdf (visited on 12/06/2020).
Palmer, Carole L., Nicholas M. Weber, and Melissa H. Cragin (2011). „The analytic potential of scientific data: Understanding re-use value“. In: *Proceedings of the American Society for Information Science and Technology* 48.1, pp. 1–10. DOI: 10.1002/meet.2011.14504801174.
Parsons, M. and R. Duerr (2006). „Designating user communities for scientific data: challenges and solutions“. In: *Data Science Journal* 4.0, pp. 31–38. DOI: 10.2481/dsj.4.31.
Parsons, M. and P. Fox (2013). „Is Data Publication the Right Metaphor?“ In: *Data Science Journal* 12, WDS32–WDS46. DOI: 10.2481/dsj.WDS-042.

Peters, Isabella et al. (2016). „Research data explored: an extended analysis of citations and altmetrics“. In: *Scientometrics* 107.2, pp. 723–744. DOI: 10.1007/s11192-016-1887-4.

Plantin, Jean-Christophe (2018). „Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science:“ in: *Science, Technology, & Human Values* 44.1, pp. 52–73. DOI: 10.1177/0162243918781268.

Pomerantz, Jeffrey (2015). *Metadata*. Cambridge, MA; London: The MIT Press.

Quarati, Alfonso and Juliana E Raffaghelli (2020). „Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case“. In: *Journal of Information Science* online first. DOI: 10.1177/0165551520961048.

Recker, Jonas and Stefan Müller (2015). „Preserving the Essence: Identifying the Significant Properties of Social Science Research Data“. In: *New Review of Information Networking* 20.1, pp. 229–235. DOI: 10.1080/13614576.2015.1110404.

Research Data Alliance FAIR Data Maturity Model Working Group (2020). „FAIR Data Maturity Model: specification and guidelines“. In: DOI: 10.15497/RDA00050.

Riley, Jenn and National Information Standards Organization (U.S.) (2017). *Understanding metadata: what is metadata, and what is it for?* URL: http://www.niso.org/publications/understanding-metadata-riley (visited on 12/06/2020).

Robinson-Garcia, Nicolas et al. (2017). „DataCite as a novel bibliometric source: Coverage, strengths and limitations“. In: *Journal of Informetrics* 11.3, pp. 841–854. DOI: 10.1016/j.joi.2017.07.003.

Rosenberg, Daniel (2013). „Data before the Fact“. In: *"Raw data" is an oxymoron*. Ed. by Lisa Gitelman. Cambridge, MA; London: The MIT Press.

Rousidis, Dimitris et al. (2014). „Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories“. In: *Information Services & Use* 34.3, pp. 279–286. DOI: 10.3233/ISU-140746.

Rowley, Jennifer (2007). „The wisdom hierarchy: representations of the DIKW hierarchy“. In: *Journal of Information Science* 33.2, pp. 163–180. DOI: 10.1177/0165551506070706.

Rücknagel, Jessika et al. (2015). „Metadata Schema for the Description of Research Data Repositories : version 3.0“. In: DOI: 10.2312/re3.008.

Schriml, Lynn M. et al. (2020). „COVID-19 pandemic reveals the peril of ignoring metadata standards“. In: *Scientific Data* 7.1. DOI: 10.1038/s41597-020-0524-5.

Schwardmann, Ulrich (2020). „Digital Objects – FAIR Digital Objects: Which Services Are Required?“ In: *Data Science Journal* 19, p. 15. DOI: 10.5334/dsj-2020-015.

Springer, Rebecca and Danielle Cooper (2020). „Data Communities: Empowering Researcher-Driven Data Sharing in the Sciences“. In: *International Journal of Digital Curation* 15.1, p. 7. DOI: 10.2218/ijdc.v15i1.695.

Star, Susan Leigh and Karen Ruhleder (1996). „Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces“. In: *Information Systems Research* 7.1, pp. 111–134. DOI: 10.1287/isre.7.1.111.

Strasser, Bruno J. and Paul N. Edwards (2017). „Big Data Is the Answer ... But What Is the Question?“ In: *Osiris* 32.1, pp. 328–345. DOI: 10.1086/694223.

Tenopir, Carol et al. (2015). „Research Data Services in Academic Libraries: Data Intensive Roles for the Future?“ In: *Journal of eScience Librarianship* 4.2. DOI: 10.7191/jeslib.2015.1085.

Weber, Tobias and Dieter Kranzlmuller (2018). „How FAIR Can you Get? Image Retrieval as a Use Case to Calculate FAIR Metrics“. In: *2018 IEEE 14th International Conference on e-Science (e-Science)*. 2018 IEEE 14th International Conference on e-Science (e-Science). Amsterdam: IEEE, pp. 114–124. DOI: 10.1109/eScience.2018.00027.

Weber, Tobias et al. (2019). „Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research“. In: *arXiv*. arXiv: 1910.09313. URL: http://arxiv.org/abs/1910.09313 (visited on 12/06/2020).

Wehrle, Dennis and Klaus Rechert (2018). „Are Research Datasets FAIR in the Long Run?“ In: *International Journal of Digital Curation* 13.1, pp. 294–305. DOI: 10.2218/ijdc.v13i1.659.

Wieczorek, John et al. (2012). „Darwin Core: An Evolving Community-Developed Biodiversity Data Standard“. In: *PLOS ONE* 7.1, e29715. DOI: 10.1371/journal.pone.0029715.

Wilkinson, Mark D. et al. (2016). „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18.

Willis, Craig, Jane Greenberg, and Hollie White (2012). „Analysis and synthesis of metadata goals for scientific data“. In: *Journal of the American Society for Information Science and Technology* 63.8, pp. 1505–1520. DOI: 10.1002/asi.22683.

Wu, Mingfang et al. (2019). „Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories“. In: *Data Science Journal* 18, p. 3. DOI: 10.5334/dsj-2019-003.

Yoon, Ayoung (2014). „End users' trust in data repositories: definition and influences on trust development“. In: *Archival Science* 14.1, pp. 17–34. DOI: 10.1007/s10502-013-9207-8.

York, Jeremy, Myron Gutmann, and Francine Berman (2018). „What Do We Know about the Stewardship Gap“. In: *Data Science Journal* 17, p. 19. DOI: 10.5334/dsj-2018-019.

Zeng, Marcia Lei and Jian Qin (2016). *Metadata*. Chicago, IL: Neal-Schuman.

Appendices

A Elements in the DataCite Metadata Schema by NISO metadata types

The table below summarizes the elements in the DataCite Metadata Schema and the corresponding NISO metadata types (Riley and National Information Standards Organization (U.S.), 2017, p. 6f).

element name	metadata type
identifier	descriptive metadata
identifierType	descriptive metadata
creatorName	descriptive metadata
nameType (creator)	descriptive metadata
givenName (creator)	descriptive metadata
familyName (creator)	descriptive metadata
nameIdentifier (creator)	descriptive metadata
nameIdentifierScheme (creator)	descriptive metadata
schemeURI (creator)	descriptive metadata
affiliation (creator)	descriptive metadata
affiliationIdentifier (creator)	descriptive metadata
affiliationIdentifierScheme (creator)	descriptive metadata
affiliationSchemeURI (creator)	descriptive metadata
title	descriptive metadata
titleType	descriptive metadata
publisher	descriptive metadata
publicationYear	descriptive metadata
resourceType	technical metadata
resourceTypeGeneral	technical metadata
subject	descriptive metadata
subjectScheme	descriptive metadata
schemeURI (subject)	descriptive metadata
valueURI (subject)	descriptive metadata
contributorType	descriptive metadata
contributorName	descriptive metadata
nameType (contributor)	descriptive metadata
givenName (contributor)	descriptive metadata
familyName (contributor)	descriptive metadata
nameIdentifier (contributor)	descriptive metadata
nameIdentifierScheme (contributor)	descriptive metadata
schemeURI (contributor)	descriptive metadata
affiliation (contributor)	descriptive metadata
affiliationIdentifier (contributor)	descriptive metadata
affiliationIdentifierScheme (contributor)	descriptive metadata
affiliationSchemeURI (contributor)	descriptive metadata
date	descriptive metadata

dateType	descriptive metadata
dateInformation	descriptive metadata
language	descriptive metadata
alternateIdentifier	structural metadata
alternateIdentifierType	structural metadata
relatedIdentifier	structural metadata
relatedIdentifierType	structural metadata
relationType	structural metadata
relatedMetadataScheme	structural metadata
schemeURI (relatedIdentifier)	structural metadata
schemeType (relatedIdentifier)	structural metadata
resourceTypeGeneral (relatedIdentifier)	structural metadata
size	technical metadata
format	technical metadata
version	structural metadata
rights	rights metadata
rightsURI	rights metadata
rightsIdentifier	rights metadata
rightsIdentifierScheme	rights metadata
schemeURI (rightsIdentifier)	rights metadata
description	descriptive metadata
descriptionType	descriptive metadata
geoLocation	descriptive metadata
geoLocationPoint	descriptive metadata
pointLongitude (geoLocationPoint)	descriptive metadata
pointLatitude (geoLocationPoint)	descriptive metadata
geoLocationBox	descriptive metadata
westBoundLongitude (geoLocationBox)	descriptive metadata
eastBoundLongitude (geoLocationBox)	descriptive metadata
southBoundLongitude (geoLocationBox)	descriptive metadata
northBoundLongitude (geoLocationBox)	descriptive metadata
geoLocationPlace	descriptive metadata
geoLocationPolygon	descriptive metadata
polygonPoint	descriptive metadata
pointLongitude (polygonPoint)	descriptive metadata
pointLatitude (polygonPoint)	descriptive metadata
inPolygonPoint	descriptive metadata
pointLongitude (inPolygonPoint)	descriptive metadata
pointLatitude (inPolygonPoint)	descriptive metadata
fundingReference	descriptive metadata
funderName	descriptive metadata
funderIdentifier	descriptive metadata
funderIdentifierType	descriptive metadata
schemeURI (funderIdentifier)	descriptive metadata
awardNumber	descriptive metadata
awardURI	descriptive metadata
awardTitle	descriptive metadata

B Repositories in the sample

re3data ID	repository name	repository type	certification status
r3d100012587	EnviDat	disciplinary	FALSE
r3d100012825	Forschungsdaten-Repositorium der LUH	institutional	FALSE
r3d100012001	Illinois Data Bank	institutional	FALSE
r3d100012330	RADAR	other	FALSE
r3d100012646	Federated Research Data Repository	other	FALSE
r3d100012505	ORDaR	disciplinary	FALSE
r3d100012064	University of Reading Research Data Archive	institutional	FALSE
r3d100012927	Data Commons	institutional	FALSE
r3d100012140	Brunel figshare	institutional	FALSE
r3d100012190	ZBW Journal Data Archive	disciplinary	FALSE
r3d100012405	Research Data at Essex	institutional	FALSE
r3d100013062	Ifsttar research data	institutional	FALSE
r3d100012157	Fairdata IDA Research Data Storage Service	other	FALSE
r3d100011601	Structural Biology Data Grid	disciplinary | institutional	FALSE
r3d100012145	melbourne.figshare.com	institutional	FALSE
r3d100012633	ZivaHub	institutional	FALSE
r3d100011864	OpenKIM	disciplinary	FALSE
r3d100011890	Ag Data Commons	disciplinary	FALSE
r3d100011945	Research Data Leeds Repository	institutional	FALSE
r3d100012414	UEL Research Repository	institutional	FALSE
r3d100012147	Stockholm University repository for data	institutional	FALSE
r3d100011947	University of Bath Research Data Archive	institutional	FALSE
r3d100012417	UCL Discovery	institutional	FALSE
r3d100012384	CaltechDATA	institutional	FALSE
r3d100012369	Code Ocean	disciplinary	FALSE
r3d100012335	GFZ Data Services	disciplinary	FALSE
r3d100010216	4TU.ResearchData | science.engineering.design	disciplinary | institutional	TRUE
r3d100012564	ScholarBank@NUS	institutional	FALSE
r3d100011662	Landcare Research Data Repository	disciplinary | institutional	FALSE
r3d100010299	World Data Center for Climate	disciplinary	TRUE
r3d100010478	GigaDB	disciplinary	FALSE
r3d100012557	ETH Zürich Research Collection	institutional	FALSE
r3d100010731	Open Data LMU	institutional | other	FALSE
r3d100011038	Qualitative Data Repository	disciplinary	TRUE
r3d100012143	Loughborough Data Repository	institutional	FALSE

r3d100012965	IFREMER-SISMER Portail de données marines	disciplinary	TRUE
r3d100000044	DRYAD	other	FALSE
r3d100012538	DataverseNO	disciplinary | institutional | other	TRUE
r3d100000006	Archaeology Data Service	disciplinary	TRUE
r3d100010066	figshare	other	FALSE
r3d100010468	Zenodo	other	FALSE
r3d100010664	World Stress Map	disciplinary	TRUE
r3d100011108	heiDATA	institutional | other	FALSE
r3d100012757	RODARE	institutional	FALSE
r3d100013029	TUdatalib	institutional	FALSE
r3d100013084	SURF Data Repository	other	FALSE
r3d100013275	GRO.data	institutional	FALSE

C Definitions of elements in the Datacite Metadata Schema

The definitions in the table below are direct quotes from the documentation (DataCite Metadata Working Group, 2019; p. 13ff).

element name	definition	added in version	obligation level
identifier	The Identifier is a unique string that identifies a resource.	2.0	mandatory
identifierType	The type of Identifier.	2.0	mandatory
creatorName	The main researchers involved in producing the data, or the authors of the publication, in priority order.	2.0	mandatory
nameType (creator)	The type of name.	4.1	optional
givenName (creator)	The personal or first name of the creator.	4.0	optional
familyName (creator)	The surname or last name of the creator.	4.0	optional
nameIdentifier (creator)	Uniquely identifies an individual or legal entity, according to various schemas.	2.0	optional
nameIdentifierScheme (creator)	The name of the name identifier schema.	2.0	optional
schemeURI (creator)	The URI of the name identifier schema.	3.0	optional
affiliation (creator)	The organizational or institutional affiliation of the creator.	3.1	optional
affiliationIdentifier (creator)	Uniquely identifies the organizational affiliation of the creator.	4.3	optional
affiliationIdentifierScheme (creator)	The name of the affiliation identifier schema.	4.3	optional
affiliationSchemeURI (creator)	The URI of the affiliation identifier schema.	4.3	optional
title	A name or title by which a resource is known.	2.0	mandatory
titleType	The type of title.	2.0	optional
publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.	2.0	mandatory
publicationYear	The year when the data was or will be made publicly available.	2.0	mandatory
resourceType	A description of the resource.	2.0	mandatory
resourceTypeGeneral	The general type of a resource.	2.0	mandatory

subject	Subject, keyword, classification code, or key phrase describing the resource.	2.0	recommended
subjectScheme	The name of the subject scheme or classification code or authority if one is used.	2.0	recommended
schemeURI (subject)	The URI of the subject identifier scheme.	3.0	optional
valueURI (subject)	The URI of the subject term.	4.0	optional
contributorType	The type of contributor of the resource.	2.0	optional
contributorName	The full name of the contributor.	2.0	recommended
nameType (contributor)	The type of name.	4.1	optional
givenName (contributor)	The personal or first name of the contributor.	4.0	optional
familyName (contributor)	The surname or last name of the contributor.	4.0	optional
nameIdentifier (contributor)	Uniquely identifies an individual or legal entity, according to various schemes.	2.0	optional
nameIdentifierScheme (contributor)	The name of the name identifier scheme.	2.0	optional
schemeURI (contributor)	The URI of the name identifier scheme.	3.0	optional
affiliation (contributor)	The organizational or institutional affiliation of the contributor.	3.1	optional
affiliationIdentifier (contributor)	Uniquely identifies the organizational affiliation of the contributor.	4.3	optional
affiliationIdentifierScheme (contributor)	Name of the affiliation identifier schema.	4.3	optional
affiliationSchemeURI (contributor)	URI of the affiliation identifier schema.	4.3	optional
date	Different dates relevant to the work.	2.0	recommended
dateType	The type of date.	2.0	recommended
dateInformation	Specific information about the date, if appropriate.	4.1	optional
language	The primary language of the resource.	2.0	recommended
alternateIdentifier	An identifier or identifiers other than the primary Identifier applied to the resource being registered.	2.0	recommended
alternateIdentifierType	The type of the AlternateIdentifier.	2.0	recommended
relatedIdentifier	Identifiers of related resources. These must be globally unique identifiers.	2.0	recommended
relatedIdentifierType	The type of the RelatedIdentifier.	2.0	recommended
relationType	Description of the relationship of the resource being registered (A) and the related resource (B).	2.0	recommended
relatedMetadataScheme	The name of the scheme.	3.0	optional
schemeURI (relatedIdentifier)	The URI of the relatedMetadataScheme.	3.0	optional
schemeType (relatedIdentifier)	The type of the relatedMetadataScheme, linked with the schemeURI.	3.0	optional

resourceTypeGeneral (relatedIdentifier)	The general type of the related resource.	4.2	optional
size	Size (e.g. bytes, pages, inches, etc.) or duration (extent), e.g. hours, minutes, days, etc., of a resource.	2.0	recommended
format	Technical format of the resource.	2.0	recommended
version	The version number of the resource.	2.0	recommended
rights	Any rights information for this resource.	2.0	recommended
rightsURI	The URI of the license.	3.0	optional
rightsIdentifier	A short, standardized version of the license name.	4.2	optional
rightsIdentifierScheme	The name of the scheme.	4.2	optional
schemeURI (rightsIdentifier)	The URI of the rightsIdentifierScheme.	4.2	optional
description	All additional information that does not fit in any of the other categories. May be used for technical information.	2.0	recommended
descriptionType	The type of the Description.	2.0	recommended
geoLocation	Spatial region or named place where the data was gathered or about which the data is focused.	3.0	recommended
geoLocationPoint	A point location in space.	3.0	recommended
pointLongitude (geoLocationPoint)	Longitudinal dimension of point.	4.0	recommended
pointLatitude (geoLocationPoint)	Latitudinal dimension of point.	4.0	recommended
geoLocationBox	The spatial limits of a box.	3.0	recommended
westBoundLongitude (geoLocationBox)	Western longitudinal dimension of box.	4.0	recommended
eastBoundLongitude (geoLocationBox)	Eastern longitudinal dimension of box.	4.0	recommended
southBoundLongitude (geoLocationBox)	Southern latitudinal dimension of box.	4.0	recommended
northBoundLongitude (geoLocationBox)	Northern latitudinal dimension of box.	4.0	recommended
geoLocationPlace	Description of a geographic location.	3.0	recommended
geoLocationPolygon	A drawn polygon area, defined by a set of points and lines connecting the points in a closed chain.	4.0	recommended
polygonPoint	A point location in a polygon.	4.0	recommended
pointLongitude (polygonPoint)	Longitudinal dimension of point.	4.0	recommended
pointLatitude (polygonPoint)	Latitudinal dimension of point.	4.0	recommended
inPolygonPoint	For any bound area that is larger than half the earth, define a (random) point inside.	4.1	recommended
pointLongitude (inPolygonPoint)	Longitudinal dimension of point.	4.1	recommended

pointLatitude (inPolygonPoint)	Latitudinal dimension of point.	4.1	recommended
fundingReference	Information about financial support (funding) for the resource being registered.	4.0	recommended
funderName	Name of the funding provider.	4.0	recommended
funderIdentifier	Uniquely identifies a funding entity, according to various types.	4.0	recommended
funderIdentifierType	The type of the funderIdentifier.	4.0	recommended
schemeURI (funderIdentifier)	The URI of the funder identifier schema.	4.3	recommended
awardNumber	The code assigned by the funder to a sponsored .award (grant).	4.0	recommended
awardURI	The URI leading to a page provided by the funder for more information about the award (grant).	4.0	recommended
awardTitle	The human readable title or name of the award .(grant).	4.0	recommended